Contents

Chapter 1

Introduction

1.0 Preliminaries

1.0.1 Motivation

Causal inference is increasingly being recognized as a crucial part of science. Understanding cause-effect relationships – rather than mere associations – is the primary goal in many if not most scientific fields (even if this goal is not clearly stated or only implied).

Here are some typical examples of important substantive causal questions that cannot be answered with associations alone:

- medicine: which cancer treatments are best for which patients?
- criminology: would more strict gun laws result in fewer homicides?
- education: how does class size impact student outcomes?

Here are some other examples of substantive questions I have personally worked on, where causal inference concepts and tools were absolutely necessary:

- Do high-tech neonatal intensive care units (NICUs) improve mortality rates for premature infants [Kennedy et al., 2019b]? Here one cannot just compare mortality rates at high-tech versus low-tech NICUs, since sicker babies are overwhelmingly more likely to be treated at high-tech NICUs. This "unadjusted" comparison would make it look like high-tech NICUs are harmful. In fact, in this case many important features capturing babies' health are missing. What should be done?
- Incarceration is a colossal industry in the United States, with over 2.3 million people currently confined in a correctional facility and at least twice that number held on probation or parole (Wagner & Rabuy 2016). What are the effects of this mass incarceration phenomenon on inmates' and families' sociological outcomes, including marriage rates [Kennedy, 2019]? This is a difficult question to study: for example, those who are incarcerated can be very different from those who are not, and incarceration status changes over time and is a result of myriad factors.

- Would decreasing nurse staffing affect hospitals' readmission rates [Kennedy et al., 2017]? On the one hand, reducing staffing might lead to unmet medical demands; on the other, it might lead to less overworked and more alert staff. Unadjusted comparisons will again be broken since hospitals differ in many important ways that could be related to both nurse staffing and excess readmissions.
- Does canvassing improve voter turnout [Kennedy et al., 2019a]? There exist several large-scale randomized experiments that were conducted to help assess this important policy question. Due to the randomization, confounding is not an issue... or is it? In fact some voters could not be contacted even though they were assigned to to be. Should they be counted in the control or treatment group?

This course will help you put these kinds of substantive questions into a clear and concise mathematical framework, exposing what assumptions are necessary to draw causal conclusions, and give you flexible tools for assessing such questions from complex data.

This will be addressed in detail shortly, but the main difference between causal questions and non-causal or associational ones can be stated succinctly as follows. Associational non-causal questions are about *how things are*; causal questions on the other hand are about *how things would have been*, if circumstances changed. Causal questions are inherently *counterfactual*.

1.0.2 What This Course Covers

The purpose of this course is to give a thorough introduction to the foundations as well as modern developments of statistical causal inference, including topics such as:

- randomized experiments
- unconfounded observational studies
- effect modification
- instrumental variables
- regression discontinuity
- mediation & interaction
- nonparametric bounds
- sensitivity analysis
- graphical models
- time-varying treatments
- dynamic & stochastic interventions
- optimal treatment regimes
- principal stratification
- interference
- matching, weighting, & regression
- nonparametric efficiency theory
- functional estimation
- heterogeneous treatment effects

Most of our discussions will follow the same basic template:

1. Clearly define the counterfactual parameter(s) of interest;
2. State and assess the assumptions necessary for identification;
3. Describe and implement various tools for estimation, and interpret results.

1.0.3 Statistical Review

Some parts of causal inference can be conveyed visually with plots and graphs, or verbally using everyday language. However, a deeper understanding requires mathematics and statistics; this will be crucial in this course since it has a special focus on the statistical aspects of causal inference.

Here I give a brief review of some crucial concepts that will be important if not necessary for large sections of the course. If any of these seem foreign to you, I encourage you to brush up; some readable and relevant textbooks for the basics include Boos and Stefanski [2013] and van der Vaart [2000].

You should have a clear understanding and recall of all the following basic concepts:

- independence, random variable, sample versus population, iid, estimation, regression, bias, variance, mean squared error, confidence intervals, hypothesis testing

Here are some fundamental results and definitions that we will rely upon extensively.

Result 1.1 (Iterated expectation)**.** Let X and Y be any two random variables. The law of iterated expectation states that

$$\mathbb{E}(Y) = \mathbb{E}\{\mathbb{E}(Y \mid X)\} = \int \mathbb{E}(Y \mid X = x)\, d\mathbb{P}(x).$$

Definition 1.1 (Big-O)**.** A sequence of random variables $\{X_n\}$ is bounded in probability, i.e., $X_n = O_{\mathbb{P}}(1)$, if for any $\epsilon > 0$ there exists $M, N < \infty$ such that

$$\mathbb{P}(|X_n| > M) < \epsilon$$

for all $n \geq N$. We say $X_n = O_{\mathbb{P}}(r_n)$ for some sequence $\{r_n\}$ if $X_n/r_n = O_{\mathbb{P}}(1)$.

Definition 1.2 (Little-O)**.** A sequence of random variables $\{X_n\}$ converges in probability to zero as $n \to \infty$, i.e., $X_n = o_{\mathbb{P}}(1)$, if for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon) = 0.$$

We say $X_n = o_{\mathbb{P}}(r_n)$ for some sequence $\{r_n\}$ if $X_n/r_n = o_{\mathbb{P}}(1)$.

Definition 1.3 (Consistency)**.** An estimator $\widehat{\psi}$ is consistent for a target quantity ψ, written

$$\widehat{\psi} \xrightarrow{p} \psi$$

if $\widehat{\psi} - \psi$ converges in probability to zero, i.e., $\widehat{\psi} - \psi = o_{\mathbb{P}}(1)$. We say $\widehat{\psi}$ is consistent at rate $r_n \to \infty$ (e.g., $r_n = \sqrt{n}$) if

$$r_n(\widehat{\psi} - \psi) = O_{\mathbb{P}}(1).$$

In this case r_n (or $1/r_n$) is called the *rate of convergence* of $\widehat{\psi}$.

Definition 1.4 (Convergence in distribution)**.** A sequence of random variables $\{X_n\}$ converges in distribution to Z, written

$$X_n \rightsquigarrow Z$$

if $\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(Z \leq x)$ at all continuity points.

Definition 1.5 (Sample average)**.** Many important estimators can be written as sample averages, at least asymptotically. We use the shorthand

$$\mathbb{P}_n\{f(Z)\} = \mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(Z_i).$$

Result 1.2 (Law of large numbers)**.** If $(X_1, ..., X_n)$ are an iid sample from $\mathbb{P}$ with mean $\mathbb{E}(X) < \infty$ then

$$\mathbb{P}_n(X) \xrightarrow{p} \mathbb{E}(X).$$

Result 1.3 (Central limit theorem)**.** If $(X_1, ..., X_n)$ are an iid sample from $\mathbb{P}$ with mean $\mathbb{E}(X) < \infty$ and variance $\text{var}(X) < \infty$ then

$$\sqrt{n}\Big\{\mathbb{P}_n(X) - \mathbb{E}(X)\Big\} \rightsquigarrow N(0, \text{var}(X)).$$

1.1 Association versus Causation

The fundamental difference between association and causation is that association concerns *how things are*, while causation concerns *how things would have been*, had something changed in the system we are observing (or had something been intervened upon). Causal inference is inherently *counterfactual*: it concerns "what might have happened if X occurred", when in fact X may have not occurred in reality.

Let's consider the examples discussed earlier. An associational question in the NICU example is:

Are mortality rates higher in high-tech or low-tech NICUs?

whereas a causal question is:

Would the infants treated at low-tech NICUs have fared better at high-tech NICUs?

An associational question in the incarceration example is:

Are marriage rates lower among those who are incarcerated longer?

whereas a causal question is:

If incarceration rates decreased, would marriage rates change?

Can you spot the differences?

Associational questions ask about how things are – they do not require us to imagine intervening upon or changing the system we are observing. Causal questions are different; they ask how things *would have been* if something fundamental had changed.

In some cases, association and causation are easy to distinguish. We know in our gut that there is an association between the number of cigarette lighters people own and their risk of lung cancer, not because lighters are deadly but because people who smoke are more likely to own more lighters.

At http://www.tylervigen.com/ you can find even more ridiculous examples of spurious correlations. My favorite is shown in Figure 1.1

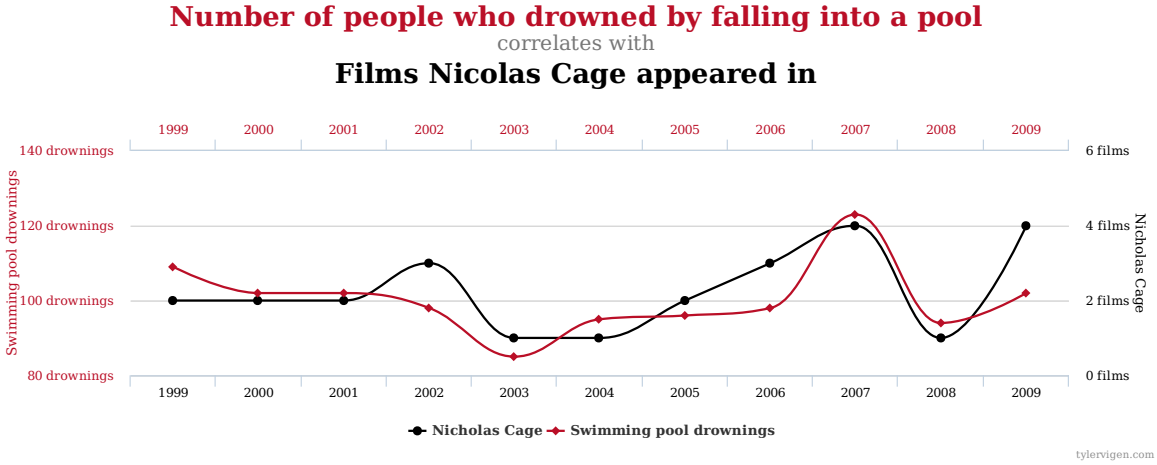


Figure 1.1: An excellent example of a spurious correlation.

However, in more typical cases the difference between association and causation can be quite subtle or hard to spot, even for experts. Many statisticians have been guilty of conflating association and causation (though they are by no means alone in this respect), whether intentionally or not. According to Wasserman (1999):

There are two types of statisticians: those who do causal inference and those who lie about it.

Here is an example of a common conflation of association and causation, which is propagated in introductory statistics courses across many prestigious universities:

Example 1.1. What is the interpretation of the coefficient β_1 in the following elementary linear regression model?

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

It is popular to say β_1 represents "the expected change in outcome Y if covariate X_1 were increased by one, keeping other covariates constant." However this is incorrect without adding extra causal assumptions. Otherwise it only represents the expected difference in outcomes for two units who *happen to have* the same covariate values $(X_2, ..., X_p)$, but whose X_1 values *happen to differ* by one.

There is a crucial distinction between the former and latter interpretations: the latter claims to say what happens when the system is *changed or intervened upon* (i.e., what would happen if we, contrary to fact, increased a covariate by one unit, while keeping all others constant), whereas the other merely says something about how the system is in reality. The former imagines changing or intervening on one group of subjects, whereas the latter compares two different groups that happen to differ, for unspecified reasons.

A good exercise is to consider which of the following concepts you would classify as associational or causal [Pearl, 2009a]:

- correlation
- regression
- dependence
- conditional independence
- likelihood
- collapsibility
- propensity score
- risk ratio
- odds ratio
- marginalization
- conditionalization
- randomization
- influence
- effect
- confounding
- "holding constant"
- spurious correlation
- instrumental variables
- intervention
- explanation
- attribution

1.2 Causal Language & Notation

Historically there has been some tension between statistics and causality; one of the reasons for this is that purely associational statistics does not have the linguistic capacity for counterfactuals.

For example suppose we observe an iid sample of $Z = (X, A, Y)$ where X are covariates, A is a treatment or exposure, and Y is an outcome. It is not possible to denote intervention on A without some new notation: a new language is needed.

There are three common ways to express counterfactual quantities:

1. structural equations
2. graphs (plus a structural model)
3. potential outcomes

These languages can all act together in concert, and in a formal sense are all equivalent. In practice I find that no one language dominates – one may be most useful in some settings, another in others.

1.2.1 Structural Equations

Structural equations began with the work of Sewall Wright in the 1920s, and are particularly popular to this day, especially among economists. Structural equations are really just usual equations that are causal by assumption.

The first structural equations were always linear and Gaussian, e.g.,

$$\begin{aligned} X &= \epsilon_X \\ A &= \alpha X + \epsilon_A \\ Y &= \beta_0 + \beta_1 X + \psi A + \epsilon_Y \end{aligned}$$

for $\epsilon_t \sim N(0, \sigma_t^2)$ error terms. Note this looks exactly like a usual linear regression model; the distinction is not in the notation, instead it is imbued with extra-notational meaning. In particular, by calling this model "structural", one is saying it represents how nature actually works.

You can think of this as an ordered computer program:

1. first, nature draws an $X \sim N(0, \sigma_X^2)$.
2. then, nature draws an $A \sim N(\alpha X, \sigma_A^2)$.
3. finally, nature draws a $Y \sim N(\beta_0 + \beta_1 X + \psi A, \sigma_Y^2)$.

Importantly, the error terms include any and all variables that influence the left-hand-side of the corresponding equation, i.e., all those factors nature uses to assign values.

The ordering and the left-hand versus right-hand-side distinctions are crucial in a structural equation model. You might be tempted to rearrange it to write

$$X = (A - \epsilon_A)/\alpha \ , \ X = \epsilon_X \ , \ Y = \beta_0 + \beta_1 X + \psi A + \epsilon_Y$$

but this loses structural meaning: nature is assigning X twice (and not assigning A).

Structural equation models were generalized to the nonparametric case in the 1990s by Pearl and Spirtes, where for example one might instead write

$$X = f_X(\epsilon_X)$$
$$A = f_A(X, \epsilon_A)$$
$$Y = f_Y(X, A, \epsilon_Y)$$

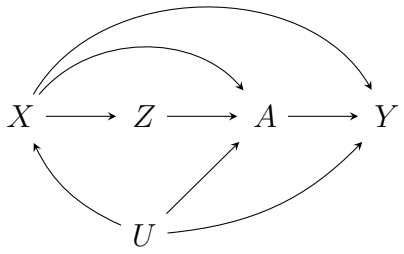
Assumptions about confounding and other structure are imposed via assumptions about errors ϵ_t. Interventions are represented by setting values within the structural equations; e.g., if we wanted to set A to a, we would write:

$$X = f_X(\epsilon_X)$$
$$A = a$$
$$Y = f_Y(X, a, \epsilon_Y)$$

A downside of structural equations is that it can be easy to confuse them with non-structural equations.

1.2.2 Graphs

Graphs are a helpful way to visualize causal structure. Here is an example of a graph:



The meaning in a graph comes from the lack of arrows – in other words, arrows mean there may or may not be a causal relationship, but no arrow means intervening *has no effect whatsoever*. For example, the above graph implies that intervening on or changing Z can change the distribution of A but not Y directly.

Graphs mean different things depending on the underlying causal model; really graphs must be paired with an explicit model, like the structural equation models from above. For example, the above graph could be paired with

$$U = f_U(\epsilon_U)$$
$$X = f_X(U, \epsilon_X)$$
$$Z = f_Z(X, \epsilon_Z)$$
$$A = f_A(U, X, Z, \epsilon_A)$$
$$Y = f_Y(U, X, A, \epsilon_Y)$$

1.2.3 Potential Outcomes

Potential outcomes are the dominant causal language in statistics, and are what we will most often use in this class (though we will also make use of graphs and structural equations). They were first used by Jerzy Neyman in 1923 to analyze agricultural experiments, were used at least conceptually in the social sciences by Tinbergen (1930) and Haavelmo (1944), and in general observational studies by Rubin (1974).

Suppose we have data on treatment A and outcome Y on people $i = 1, ..., n$. The **potential outcome** we *would have observed* for person i had they received treatment $A = 1$ is denoted $Y_i^{a=1}$, and the potential outcome had they received control $A = 0$ is denoted $Y_i^{a=0}$. Often we will drop the i subscript, and if it is clear what is being intervened upon we will just write Y^1 or Y^0.

Note that Y represents what we actually observed, whereas Y^a represents what we *would have observed*. This is a big difference! You can imagine Y^a for different values of a representing different outcomes that would have existed in parallel universes where everything else is the same except for the choice/intervention on A.

Remark 1.1. We will (mostly) use superscripts as in Y^a to denote potential outcomes, but sometimes authors use subscripts Y_a or parentheses $Y(a)$.

Consider a simple example. Suppose I had a headache, took aspirin and subsequently found that my headache went away. Then my data could be expressed as $(A, Y) = (1, 0)$ where:

- $A = 1$ indicates that I took aspirin
- $Y = 0$ indicates that I didn't have a headache after

And here:

- Y^0 is whether I would have had a headache had I not taken aspirin
- Y^1 is whether I would have had a headache had I taken aspirin

In this case we would often imagine that I actually observed my Y^1 potential outcome, because I actually took aspirin (however, later on we will talk about when this seemingly tautological result may not hold). Note however that I certainly would not be able to observe my Y^0 potential outcome, i.e., whether my headache would have gone away had I not taken aspirin. This only exists in the parallel universe where I did not take aspirin, which I was cut off access to once I decided to take aspirin.

In fact we essentially *never* get to observe all the possible potential outcomes. Typically we only observe one: the potential outcome we would have observed under the actual circumstances (since the actual circumstances really did happen!). As mentioned above, you can imagine parallel universes representing all your potential outcomes, but

once you take a given treatment or engage in a particular policy, the paths through these universe fork off and you can only access the one you actually are in. Holland [1986] called this the "fundamental problem of causal inference": we want to learn about potential outcomes, but only see outcomes from the actual world, not the counterfactual ones in parallel universes. For this reason you can often view causal inference as a big missing data problem.

Remark 1.2. Although we have started our discussion with binary treatments, there is conceptually no difficulty in moving to multiple or continuous treatments. We can simply write $Y^{a_1,...,a_T}$ for the potential outcome had we intervened and set treatments or exposures to $(A_1, ..., A_T) = (a_1, ..., a_T)$, or imagine Y^a as a curve in a.

1.3 Causal Effects

In principle one can conceptualize an individual causal effect $Y_i^{a=1} - Y_i^{a=0}$ for a particular unit i. For example, suppose A is an indicator for whether aspirin was taken and Y is a headache indicator:

- $Y_i^{a=1} - Y_i^{a=0} = 0$ means the subject is either doomed or immune: $(Y_i^{a=1}, Y_i^{a=0}) = (1, 1)$ or $(Y_i^{a=1}, Y_i^{a=0}) = (0, 0)$
- $Y_i^{a=1} - Y_i^{a=0} = -1$ means the subject was saved: $(Y_i^{a=1}, Y_i^{a=0}) = (0, 1)$
- $Y_i^{a=1} - Y_i^{a=0} = 1$ means the subject was harmed: $(Y_i^{a=1}, Y_i^{a=0}) = (1, 0)$

However, because of the fundamental problem of causal inference, the quantity $Y_i^{a=1} - Y_i^{a=0}$ (or any other unit-level quantity depending on multiple potential outcomes) cannot usually be observed. For example, if investigators assigned treatment and then measured an outcome, and subsequently assigned control and measured the outcome, then for those outcomes to represent true counterfactuals, one would have to assume the outcomes would have been the same if measured at exactly the same time under otherwise identical circumstances (e.g., no carry-over effects, etc.). These kinds of assumptions are typically too strong to employ in practice.

Although the fundamental problem of causal inference makes it sound like causal effects are an impossible target, we will see that in some studies one can actually accurately estimate population-level average effects, for example, such as

$$\mathbb{E}(Y^{a=1} - Y^{a=0}).$$

This quantity is called the *average treatment effect* (ATE), and is probably the most popular target effect in causal inference. We will consider its estimation in a wide variety of settings (experiments, unconfounded observational studies, studies with unmeasured confounding, etc.). In words, the ATE represents the mean outcome we would have observed in a population if *all* versus *none* were treated.

One might wonder what precisely the expectation is over in the ATE parameter. There are a few ways to think about this. The first is that there exists some finite sample of n subjects, and then the expectation is just the average in the sample

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i^{a=1} - Y_i^{a=0}).$$

A second approach is to view data on n subjects not as the entire population itself, but instead as a sample from a larger population, sometimes viewed as so large it can be treated as infinite (called a superpopulation). Then the expectation would represent the expectation in the superpopulation. This is the viewpoint often taken in statistics where one generalizes from a sample to learn about a much larger population of interest. A mathematically equivalent way to think about the latter setup is to suppose the potential outcomes Y^a are random variables generated independently from a given (joint) distribution. Typically all these interpretations yield the same or very similar methods: when sampling without replacement from a finite population, the error in the superpopulation approach will be small as long as the population is large. Finite-sample results often require some tedious calculations, while the superpopulation setup is arguably more clean and clear, while still preserving most of the important ideas; thus we will tend to use the latter in this course.

The ATE is by no means the only causal parameter of substantive or theoretical interest. One can also consider:

- other summaries such as risk or odds ratios: $\frac{\mathbb{P}(Y^{a=1}=1)/\mathbb{P}(Y^{a=1}=0)}{\mathbb{P}(Y^{a=0}=1)/\mathbb{P}(Y^{a=0}=0)}$
- distributional effects: $\mathbb{P}(Y^{a=1} \leq y)$
- conditional effects: $\mathbb{E}(Y^{a=1} - Y^{a=0} \mid V = v)$
- effects of joint or multiple treatments: $\mathbb{E}(Y^{m,a})$ or $\mathbb{E}(Y^{a_1,\ldots,a_T})$
- effects of dynamic or stochastic interventions $\mathbb{E}(Y^Q)$ for Q an intervention that is random and/or depends on other variables
- optimal treatment regimes: $\arg\max_d \mathbb{E}(Y^{d(X)})$ for $d : \mathcal{X} \mapsto \mathcal{A}$ a treatment rule

along with countless other variations. One of the most exciting parts of causal inference is proposing a new and unusual variant of a causal effect based on some substantive problem of interest.

Remark 1.3. Some authors make a point of requiring causal effects to be a contrast, like the ATE; however we will consider any counterfactual estimand a causal effect.

1.4 Identification

Most causal inference problems consist of three crucial parts:

1. choosing a target parameter
2. identification (or lack thereof)
3. estimation and inference

This trilogy will play a huge role in the course, and you will see it repeated often and throughout.

Typically, the choice of target parameter should depend on the scientific question: What kind of intervention is of interest? Treating everyone versus no one? What outcome measure matters? Is it the mean outcome? A quantile? Are population-wide or subgroup effects of interest?

However, in practice the target parameter is often only defined vaguely (e.g., as "the effect") or is chosen based on convenience (e.g., a coefficient in a likely misspecified and somewhat arbitrary logistic regression model). I have encountered two cultures in applied statistics:

1. Model the entire data generating process, and then use that model to answer any and all scientific questions.
2. Start with a specific research question, and tailor the analysis and estimation procedure accordingly.

I am a big fan of the second approach: it forces one to think hard about the science and the particular goal, and further a one-size-fits-all model is often provably not optimal for all questions (i.e., better statistical properties can be achieved by tailoring).

One way to pick a target parameter is to ask: what experiment would you have conducted if there were no ethical or feasibility concerns, and the universe was at your control? For example:

- force everyone to contribute lab values
- give everyone treatment, then go back in time and withhold treatment
- force everyone to become obese, then assess outcomes after 30 years

Counterfactual causal language such as potential outcomes lets us express these kinds of hypothetical interventions mathematically.

The next step is identification, which means expressing the causal parameter in terms of an observed data distribution.

Definition 1.6. A parameter $\psi = \psi(\mathbb{P})$ is identified if $\psi(\mathbb{P}) \neq \psi(\mathbb{Q}) \implies \mathbb{P} \neq \mathbb{Q}$ for all $\mathbb{P}$ and $\mathbb{Q}$ in the model.

In causal inference problems we typically have parameters defined on counterfactual distributions $\mathbb{P}^*$, which yield observational distributions $\mathbb{P}$ via some coarsening procedure $\mathbb{P} = f(\mathbb{P}^*)$. For example, the counterfactuals (Y^1, Y^0) have some joint distribution $\mathbb{P}^*$ in the population, but the observed outcome $Y = AY^1 + (1-A)Y^0$ only depends partially on $\mathbb{P}^*$ through the distributions $p(Y^1 \mid A = 1) = p(Y \mid A = 1)$ and $p(Y^0 \mid A = 0) = p(Y \mid A = 0)$.

A schematic for identification is given in Figure 1.2.

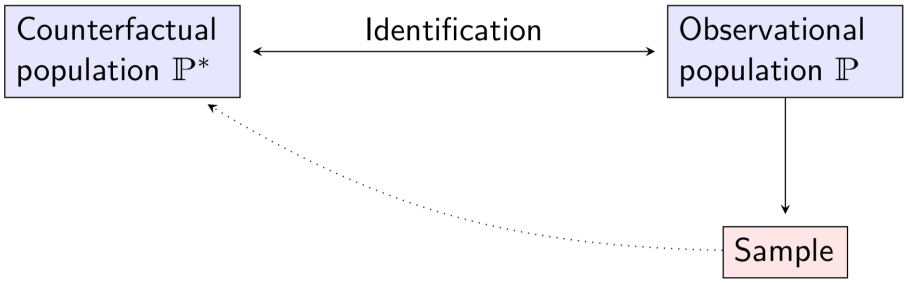


Figure 1.2: An illustration of how identification links counterfactual and observational distributions, allowing estimation of causal quantities from observational data.

Example 1.2. Suppose we observe an iid sample $(Z_1, ..., Z_n)$ with $Z = (A, Y)$. Assume $Y = AY^1 + (1-A)Y^0$ as in the aspirin example (aside: can you think of when this might be violated?). Is $\mathbb{E}(Y^1)$ identified?

Intuitively, the answer should be no since we only observe Y^1 among those with $A = 1$. To prove non-identifiability we need to construct two counterfactual distributions $\mathbb{P}^*$ and $\mathbb{Q}^*$ for which $\psi(\mathbb{P}^*) \neq \psi(\mathbb{Q}^*)$ but for which the corresponding observed data distributions $\mathbb{P}$ and $\mathbb{Q}$ are equivalent.

Let $\mathbb{P}(A = 1) = 1/2$, and let $\mathbb{P}^*(Y^a = 1 \mid A = a) = \mathbb{Q}^*(Y^a = 1 \mid A = a) = 1/2$. This means half the population is treated, and among those who are treated half would have the outcome if treated (and similarly among those who are untreated, half would have the outcome if untreated). In this case the observational distributions $\mathbb{P}$ and $\mathbb{Q}$ are the same since

$$\mathbb{P}(Y = 1 \mid A = a) = \mathbb{P}^*(Y^a = 1 \mid A = a) = \mathbb{Q}^*(Y^a = 1 \mid A = a)$$

but by the law of total expectation

$$\mathbb{E}(Y^1) = 0.25 + 0.5\mathbb{E}(Y^1 \mid A = 0).$$

Therefore if we set $\mathbb{E}_{\mathbb{P}^*}(Y^1 \mid A = 0) \neq \mathbb{E}_{\mathbb{Q}^*}(Y^1 \mid A = 0)$ we will have $\mathbb{E}_{\mathbb{P}^*}(Y^1) \neq \mathbb{E}_{\mathbb{Q}^*}(Y^1)$, so the counterfactual $\mathbb{E}(Y^1)$ is not identified. The intuition is: since we never see Y^1 for the untreated, we can vary this for $\mathbb{P}^*$ and $\mathbb{Q}^*$ without varying the observational distributions.

We saw above that, if the same observational distribution can lead to different parameter values, then $\psi = \mathbb{E}(Y^1)$ is not identified. This means that even if we knew the observational distribution completely without error, we still would not know the target parameter ψ. Before long we will discuss how to deal with non-identified parameters, e.g., by estimating bounds and doing sensitivity analyses.

Remark 1.4. It will be crucial for this course to keep in mind the sequencing:

$$\text{parameter definition} \rightarrow \text{identification} \rightarrow \text{estimation}$$

These are three essentially separate tasks, which require different tools and bring different difficulties, depending on the causal problem at hand.

Chapter 2

Simple Randomized Experiments

2.1 Why Randomization?

Suppose we observe outcomes $(Y_1, ..., Y_n)$ for n subjects, each of whom are either treated $(A = 1)$ or not $(A = 0)$, and we want to learn the effect of the treatment A on outcome Y, say on average. An initial idea might be to compare the average outcome for those n_1 who receive treatment versus the n_0 who receive control:

$$\frac{1}{n_1} \sum_{i:A_i=1} Y_i \quad \text{versus} \quad \frac{1}{n_0} \sum_{i:A_i=0} Y_i$$

However, as discussed previously, any differences we see could be spurious, i.e., explained by something else. For example, high-tech NICUs have higher mortality rates than low-tech NICUs, but this is because high-tech NICUs see the sickest infants. In general, differences in outcomes might just be due to the people receiving treatment being inherently different from those receiving control.

A second option might be to try to measure any and all variables X that could explain any differences in outcomes, and then do a stratified (adjusted) analysis that only compares subjects with the same X values. In the NICU example, one might try to measure every possible facet of babies' health, such as birthweight, gestational age, family medical history, mother's smoking status, relevant biomarker values and lab results, and so on. There are at least three severe difficulties with this approach:

1. We often simply do not know every single $X = (X_1, X_2, ..., X_{1000}, ...)$ that could explain *any* differences in outcomes.

2. Even if we did know every single possible X with certainty, it might be impossible or too expensive to measure every single one of them.

3. And even if we could measure every single X, there may be so many that the curse of dimensionality would make estimation impossibly difficult (e.g., few if any subjects would have the same or similar Xs in every dimension).

So is assumption-free causal inference hopeless? Luckily not; it turns out there is a simple yet beautiful solution if we can control who gets treatment: assign it randomly! For example, one could flip a coin to decide whether each subject gets treatment versus control. Surprisingly, the benefits of randomization were largely unknown until relatively recently in the long history of science (according to the OED its first recorded use was due to R.A. Fisher in 1926).

Why does random treatment assignment work? Randomization ensures that the treatment is completely independent of *all* subject characteristics, whether measured or not. In other words, the treated look *exactly the same* as the untreated, in expectation, and not only for all measured variables X but also for *any* unmeasured variables U. Thus any observed differences in the outcomes for the treated versus untreated must be due to the treatment, since it is the only systematic way in which the groups differ.

Another way to think about why randomization works is in terms of potential outcomes. Suppose each subject has two potential outcomes, Y^1 and Y^0, with the former revealed by treatment and the latter revealed by control, so that $Y = AY^1 + (1-A)Y^0$. Then, by randomly assigning treatment A, we are taking two random samples – one of the Y^1 values and another of the Y^0 values. Random samples yield unbiased estimators of population means, so the average outcomes in the two groups will be unbiased estimates of the corresponding average potential outcomes.

Of course, we can also prove randomization works mathematically:

Proposition 2.1. *Let* $(A, Y) \sim \mathbb{P}$ *and assume:*

1. *Consistency:* $Y = Y^a$ *whenever* $A = a$.
2. *Randomization:* $A \perp\!\!\!\perp Y^a$ *for each* a.

Then

$$\mathbb{E}(Y \mid A = a) = \mathbb{E}(Y^a).$$

Proof. It follows that

$$\mathbb{E}(Y \mid A = a) = \mathbb{E}(Y^a \mid A = a) = \mathbb{E}(Y^a)$$

using consistency in the first equality and randomization in the second. □

Remark 2.1. Make sure not to confuse $A \perp\!\!\!\perp Y^a$ with $A \perp\!\!\!\perp Y$: these are very different. $A \perp\!\!\!\perp Y^a$ means treatment is independent of potential outcomes (which can be viewed as "pre-treatment" variables that exist just prior to the treatment assignment), and reflects that treatment is not confounded; $A \perp\!\!\!\perp Y$ means treatment is independent of the *observed* outcome, and would for example be a consequence of treatment not only being unconfounded but also ineffective (e.g., $Y^1 = Y^0$). Always remember to distinguish potential outcomes from observed outcomes.

Remark 2.2. Although Proposition 1 gives an identification result for the mean potential outcome, its assumptions are sufficient for identifying the entire distribution of potential outcomes as $\mathbb{P}(Y^a \leq t) = \mathbb{P}(Y \leq t \mid A = a)$.

Proposition 1 shows that treatment assignment need not necessarily be a subject-specific coin flip – for the purposes of achieving identification of the potential outcome distribution, treatment just needs to be independent of potential outcomes. This leads to the following definition of a randomized experiment:

Definition 2.1. A study is a randomized experiment if the treatment assignment is both *probabilistic* and *known*.

There are many types of experimental designs. For example, letting $A^n = (A_1, ..., A_n)$:

- Completely randomized: n_1 of n subjects randomly assigned to treatment, i.e., $\mathbb{P}(A^n = a^n) = 1/\binom{n}{n_1}$ for $\sum_i a_i = n_1$.
- Bernoulli: Treatments assigned via independent coin flips, i.e., $\mathbb{P}(A^n = a^n) = (1/2)^n$ for every $a^n = (a_1, ..., a_n) \in \{0, 1\}^n$.
- Stratified Bernoulli: Treatments assigned via independent *biased* coin flips depending on covariates, i.e., $\mathbb{P}(A^n = a^n \mid X^n) = \prod_i \mathbb{P}(A_i = a_i \mid X_i)$.
- Matched pairs: Matched pairs are constructed and one is treated in each, i.e., $\mathbb{P}(A^n = a^n \mid X^n) = 1/2^{n/2}$.

One design may be favored over another due to efficiency or feasibility, for example.

2.2 Testing: Fisher's Sharp Null

Jerzy Neyman was the first to introduce potential outcomes (in 1923), but R.A. Fisher was the first to really advocate for randomization (in 1925).

Fisher was interested in testing the *sharp null hypothesis*

$$H_0 : Y_i^1 = Y_i^0 \text{ for all } i$$

which says that treatment has no effect whatsoever – not only is the mean of Y^1 exactly equal to that of Y^0, but the distributions are equal and further each individual potential outcome is exactly the same under both treatment and control. This is a strong null with lots of structure, in line with Fisher's perspective that one should "make your theories elaborate".

Recall to test a generic null hypothesis H_0 we need (1) a statistic T, and (2) its distribution under the null. Then one can obtain the infamous statistic known as a p-value, i.e., $\mathbb{P}_{H_0}(T \geq t_{obs})$, the chance under the null of seeing data as extreme as that which was actually observed.

To test Fisher's sharp null, we can use as a statistic any summary measure of how treatment changes outcomes; for example, a simple yet common choice is the absolute difference-in-means

$$T(A^n, Y^n) = \left| \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i \right| = \left| \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)} \right|$$

Note this test statistic will be large if the treated versus untreated means differ, but not if the treatment only changes non-central aspects of the distribution, e.g., the variance.

Armed with a test statistic, we now need to know its distribution under the null. This is actually easy, and typically part of the motivation for using the sharp null: it yields tractable null distributions, which can be computed in a non-asymptotic and distribution-free manner. To illustrate, consider the following completely randomized experiment, simulated in R:

```
> set.seed(100)
> ## simulate fake data
> n <- 10; a <- rep(c(1,0),5); y <- a*rnorm(n,1)+(1-a)*rnorm(n,-1)
> cbind(a,y)
      a          y
 [1,] 1  0.4978076
 [2,] 0 -0.9037255
 [3,] 1  0.9210829
 [4,] 0 -0.2601595
 [5,] 1  1.1169713
 [6,] 0 -1.0293167
 [7,] 1  0.4182093
 [8,] 0 -0.4891437
 [9,] 1  0.1747406
[10,] 0  1.3102968
>
> ## compute test statistic
> (tobs <- abs(mean(y[a==1])-mean(y[a==0])))
[1] 0.9001721
```

Here the observed value of the difference-in-means test statistic is $T(A^n, Y^n) \approx 0.9$. We can also compute the value of this statistic under the null, for any randomization, since under the null the potential outcomes are exactly the same, i.e., $Y^0 = Y^1 = Y$. Therefore we can obtain the null distribution of T by permuting the A^n vector (according to the known treatment assignment mechanism), while keeping the Y^n vector fixed, computing the corresponding value of the test statistic T, which yields the corresponding distribution $\mathbb{P}_{H_0}(T \leq t)$. A p-value can be computed by simply counting the proportion of permutations with test statistics larger than that which was observed.

This can be accomplished in R with:

```
> ## permute treatments to simulate null
> t <- NULL; for (j in 1:10000){
+     asim <- sample(a)
+     t <- c(t, abs(mean(y[asim==1])-mean(y[asim==0]))) }
>
> ## compute p-value
> mean(t>=tobs)
[1] 0.0837
```

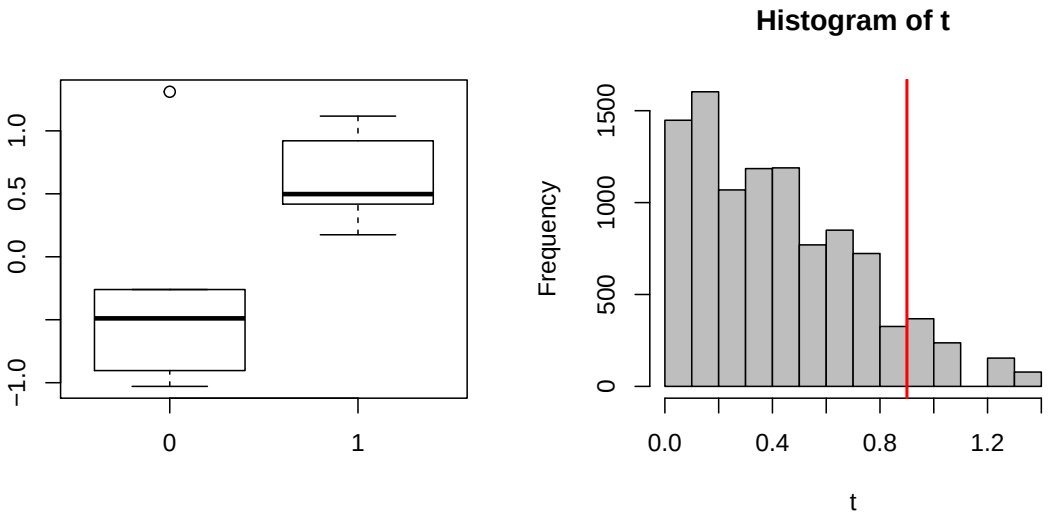


Figure 2.1: Boxplot of simulated data (left), and histogram of permutation-based null distribution with the red vertical line denoting the observed test statistic value (right).

The data and results are shown in Figure 2.1. In this simulation, the p-value is 0.084 and so there is sufficient evidence to reject the sharp null hypothesis of no individual treatment effect at level $\alpha = 0.10$.

Mathematically, for a completely randomized experiment where a fixed number n_1 are treated, the null distribution can be written as

$$\begin{aligned}\mathbb{P}_{H_0}(T \geq t) = \mathbb{P}_{H_0}\{T(A^n, y^n) \geq t\} &= \sum_{a^n \in \mathcal{A}} \mathbb{1}\{T(a^n, y^n) \geq t\}\mathbb{P}(A^n = a^n) \\ &= \sum_{a^n : \sum_i a_i = n_1} \frac{\mathbb{1}\{T(a^n, y^n) \geq t\}}{\binom{n}{n_1}}\end{aligned}$$

In theory we can compute this distribution exactly; in practice if n is large we may need to resort to simulation (e.g., sample K of the $\binom{n}{n_1}$ randomizations). However the distribution can be simulated with arbitrarily high accuracy by taking K large enough.

Remark 2.3. The null distribution calculation above treats the (potential) outcomes y^n as fixed; this can be viewed as an assumption that Y^a is not a random variable, or the probability can just be defined conditionally given the random potential outcomes.

Fisher's permutation-style test is simple but impressive: it gives an exact distribution-free p-value for testing H_0, which is valid for any n. Nonetheless here are some caveats:

- The power of the test depends heavily on the choice of statistic, e.g., the difference-in-means test statistic will have no power against a treatment that makes outcomes bimodal or otherwise more variable.
- Fisher's test is of the *sharp null* of no individual effect, not of no average effect – in fact rejecting Fisher's null could still mean there is no effect on average.

2.3 Estimation: Sample Average Effects

In the 1920s and 1930s, Fisher and Neyman had some heated debates about whether testing Fisher's sharp null should be the primary goal or not; in contrast to Fisher, Neyman advocated for estimation rather than testing, and focused on average effects. Average effects might be considered more relevant for policy decisions, since they indicate how a population would fare on average if all versus none were treated; in contrast rejecting the sharp null only indicates that treatment has *some* effect, without saying much about what kind.

The *sample average treatment effect* is given by

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0)$$

This parameter is different from those we will study later in that it is a functional of the particular sample rather than of a population distribution (i.e., strictly speaking it is a data-dependent parameter, which is why we index it with n).

Remark 2.4. In this section we treat potential outcomes as fixed, not random; this is equivalent to treating probability statements as conditional on the potential outcomes. Note however that even if the potential outcomes are fixed, the observed outcome is random since it is a function of the random treatment: $Y = Ay^1 + (1-A)y^0$.

A natural estimator for ψ_n is the difference-in-means

$$\widehat{\psi} = \frac{1}{n_1}\sum_{i:A_i=1} Y_i - \frac{1}{n_0}\sum_{i:A_i=0} Y_i = \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)}$$

We will now characterize the bias and variance of this estimator and discuss inference.

Proposition 2.2. *The difference-in-means estimator is unbiased for ψ_n in a completely randomized experiment, assuming consistency (i.e., $Y = Ay^1 + (1-A)y^0$).*

Proof. By definition, in a completely randomized experiment, we have

$$\begin{aligned}\mathbb{P}(A_1 = 1) &= \sum_{\sum_{i>1} a_i = n_1 - 1} \mathbb{P}(A_1 = 1, A_2 = a_2, ..., A_n = a_n) \\ &= \sum_{\sum_{i>1} a_i = n_1 - 1} \binom{n}{n_1}^{-1} = \frac{\binom{n-1}{n_1-1}}{\binom{n}{n_1}} = \frac{n_1}{n}\end{aligned}$$

and similarly for all other $i > 1$. Therefore

$$\begin{aligned}\mathbb{E}(\widehat{\psi}) &= \mathbb{E}\left\{\frac{1}{n_1}\sum_{i=1}^{n} A_i y_i^1 - \frac{1}{n_0}\sum_{i=1}^{n}(1 - A_i)y_i^0\right\} \\ &= \frac{1}{n_1}\sum_{i=1}^{n}\mathbb{E}(A_i)y_i^1 - \frac{1}{n_0}\sum_{i=1}^{n}\{1 - \mathbb{E}(A_i)\}y_i^0 = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0)\end{aligned}$$

where the first equality follows by consistency, and the last since $\mathbb{P}(A_i = 1) = n_1/n$. □

As mentioned previously, the intuition behind unbiasedness in this setup is that the treatments pick out random samples of Y^1 and Y^0 potential outcomes.

Now we will explore the variance of $\widehat{\psi}$, which is critical for confidence intervals and hypothesis tests; its calculation requires some care since the A_is are not independent (e.g., in the $n = 2$ case, if $A_1 = 1$ then it must be the case that $A_2 = 0$).

Proposition 2.3. *For a completely randomized experiment, and assuming consistency, the variance of the difference-in-means estimator is given by*

$$var(\widehat{\psi}) = \frac{\sigma_n^2(y^1)}{n_1} + \frac{\sigma_n^2(y^0)}{n_0} - \frac{\sigma_n^2(y^1 - y^0)}{n} \tag{2.1}$$

where

$$\sigma_n^2(v) = \frac{1}{n-1}\sum_{i=1}^{n}\left(v_i - \frac{1}{n}\sum_{j=1}^{n} v_j\right)^2$$

denotes the finite sample variance of $(v_1, ..., v_n)$.

Proof. See the appendix of Chapter 6 in Imbens and Rubin [2015]. □

A finite-sample central limit theorem implies under some regularity conditions that

$$\frac{\widehat{\psi} - \psi_n}{\sqrt{\text{var}(\widehat{\psi})}} \rightsquigarrow N(0, 1)$$

Therefore to construct large-sample confidence intervals, one needs to estimate the variance $\text{var}(\widehat{\psi})$ in (2.1). The first two terms in this variance can be estimated with

$$\widehat{\sigma}_n^2(y^a) = \frac{1}{n_a - 1} \sum_{i:A_i=a} \left(Y_i - \frac{1}{n_a} \sum_{j:A_j=a} Y_i \right)^2$$

but the third term is the finite-sample variance of the treatment effects $(y_i^1 - y_i^0)$, and involves product terms like $y_i^1 y_i^0$ which can never be observed together. Thus the third term cannot be consistently estimated; however it can be upper bounded as, for example

$$\text{var}(\widehat{\psi}) \leq \frac{\sigma_n^2(y^1)}{n_1} + \frac{\sigma_n^2(y^0)}{n_0} \tag{2.2}$$

which will yield conservative (at worst) inference, when used to construct confidence intervals. Tighter bounds can be achieved with the Cauchy-Schwarz inequality or Frechet-Hoeffding bounds [Aronow et al., 2014].

2.4 Population Average Effects

In this section we move to population rather than finite-sample effects. These effects can be especially useful for at least three reasons:

- Population effects are often of particular substantive interest: typically we might view our sample as haphazard and not particularly special, except insofar as they tell us something about some larger population from which they were drawn.
- Often population effect methods also apply without modification to sample effects, while the converse is not necessarily true; thus by studying population effects we can kill two birds with one stone. This will be discussed in detail shortly.
- Population effects can be simpler to study, easing the theoretical analyses without losing much if at all in terms of main ideas.

So here we suppose we observe an iid sample $(Z_1, ..., Z_n)$ from population distribution $\mathbb{P}$ with $Z = (A, Y)$. In this section our goal is to estimate the population average effect

$$\psi = \mathbb{E}(Y^1 - Y^0) = \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0)$$

rather than the sample average effect ψ_n from before. We rely on the following three assumptions:

1. Consistency: $Y = Y^a$ if $A = a$.
2. Bernoulli randomization: $A \perp\!\!\!\perp Y^a$ with $\mathbb{P}(A = 1) = \pi$.
3. Finite variance: Y has finite conditional variance given $A = a$.

Remark 2.5. In a Bernoulli trial, the number treated $N_1 = \sum_i A_i \sim \text{Bin}(n, \pi)$ is random, not fixed. In this section we also view the potential outcomes as random, not fixed.

Recall the difference-in-means estimator is given by

$$\widehat{\psi} = \frac{1}{\sum_i A_i} \sum_{i:A_i=1} Y_i - \frac{1}{\sum_i (1-A_i)} \sum_{i:A_i=0} Y_i = \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)}$$

Now we will study the properties of $\widehat{\psi}$: bias, variance, and limiting distribution. We will see that very precise estimation and inference are possible for the causal effect ψ in Bernoulli trials, under essentially no assumptions beyond consistency.

2.4.1 Properties of the Difference-in-Means Estimator

Theorem 2.1. *Assume consistency. In a Bernoulli trial, the difference-in-means estimator is unbiased for ψ and has variance no greater than*

$$\frac{2}{(n+1)} \left(\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi} \right)$$

where $\sigma_a^2 = var(Y \mid A = a)$.

Proof. Let $\widehat{\pi} = \mathbb{P}_n(A)$ and just consider the first term $\widehat{\mu}_1 = \mathbb{P}_n(AY)/\widehat{\pi}$ as an estimator of $\mu_1 = \mathbb{E}(Y \mid A = 1)$. We have

$$\begin{aligned} \mathbb{E}(\widehat{\mu}_1 \mid A^n) &= \frac{1}{\widehat{\pi}} \mathbb{E}\Big\{ \mathbb{P}_n(AY) \mid A^n \Big\} = \frac{1}{\widehat{\pi}} \mathbb{P}_n\Big\{ A\mathbb{E}(Y \mid A^n) \Big\} \\ &= \frac{1}{\widehat{\pi}} \mathbb{P}_n\Big\{ A\mathbb{E}(Y \mid A = 1) \Big\} = (\widehat{\pi}\mu_1)/\widehat{\pi} = \mu_1 \end{aligned}$$

where the third equality used the iid assumption. (Note: why did we not have to use randomization here? Or where did we implicitly use it?). Unbiasedness now follows by iterated expectation, and consistency follows from the weak law of large numbers and continuous mapping theorem. The logic is the same for $\widehat{\mu}_0 = \mathbb{P}_n\{(1-A)Y\}/(1-\widehat{\pi})$.

By the law of total variance we have

$$\text{var}(\widehat{\mu}_1) = \text{var}\Big\{ \mathbb{E}(\widehat{\mu}_1 \mid A^n) \Big\} + \mathbb{E}\Big\{ \text{var}(\widehat{\mu}_1 \mid A^n) \Big\}$$

Note $\text{var}\{\mathbb{E}(\widehat{\mu}_1 \mid A^n)\} = \text{var}(\mu_1) = 0$ from above, and

$$\begin{aligned} \text{var}(\widehat{\mu}_1 \mid A^n) &= \left(\frac{1}{n\widehat{\pi}} \right)^2 \sum_{i=1}^n A_i \text{var}\left(Y_i \mid A^n \right) \\ &= \left(\frac{1}{n\widehat{\pi}} \right)^2 \sum_{i=1}^n A_i \sigma_1^2 = \frac{\sigma_1^2}{N_1} \mathbb{1}(N_1 > 0) \end{aligned}$$

where we used independence and defined $\sigma_1^2 = \text{var}(Y \mid A = 1)$ and $N_1 = n\widehat{\pi}$. Now

$$\begin{aligned}\text{var}(\widehat{\mu}_1) &= \mathbb{E}\Big\{\text{var}(\widehat{\mu}_1 \mid A^n)\Big\} \\ &\leq \frac{2\sigma_1^2}{(n+1)\pi}\end{aligned}$$

by the expected binomial reciprocal result (Lemma A.2) of Devroye et al. [1996]. The same logic applies to $\widehat{\mu}_0$, and iterated expectation shows that the covariance term $\text{cov}(\widehat{\mu}_1, \widehat{\mu}_0)$ is exactly zero, which gives the result. □

Theorem 2.2. *Assume consistency. For a Bernoulli trial, the difference-in-means estimator is root-n consistent and asymptotically normal with*

$$\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N\left(0,\ \frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}\right)$$

where $\sigma_a^2 = var(Y \mid A = a)$.

Proof. We again focus on μ_1 and its estimator. Note we have

$$\begin{aligned}\widehat{\mu}_1 - \mu_1 &= \frac{\mathbb{P}_n(AY)}{\widehat{\pi}} - \mu_1 = \mathbb{P}_n\left\{\frac{A}{\widehat{\pi}}(Y - \mu_1)\right\} \\ &= \mathbb{P}_n\left\{\frac{A}{\pi}(Y - \mu_1)\right\} + \left(\frac{1}{\widehat{\pi}} - \frac{1}{\pi}\right)\mathbb{P}_n\left\{A(Y - \mu_1)\right\} \\ &= \mathbb{P}_n\left\{\frac{A}{\pi}(Y - \mu_1)\right\} + O_{\mathbb{P}}(1/\sqrt{n})O_{\mathbb{P}}(1/\sqrt{n})\end{aligned}$$

where the last equality follows by the central limit theorem, which implies $\sqrt{n}\{\mathbb{P}_n(V) - \mathbb{E}(V)\} = O_{\mathbb{P}}(1)$ for any iid V with finite mean and variance, together with the fact that $(\widehat{\pi}, \pi)$ are bounded away from zero.

Therefore

$$\widehat{\mu}_1 - \mu_1 = \mathbb{P}_n\left\{\frac{A}{\pi}(Y - \mu_1)\right\} + o_{\mathbb{P}}(1/\sqrt{n})$$

since $O_{\mathbb{P}}(1/\sqrt{n})O_{\mathbb{P}}(1/\sqrt{n}) = O_{\mathbb{P}}(1/n) = o_{\mathbb{P}}(1/\sqrt{n})$, which from the central limit theorem (together with Slutsky's theorem) gives

$$\sqrt{n}\Big(\widehat{\mu}_1 - \mu_1\Big) \rightsquigarrow N\left(0, \text{var}\left\{\frac{A}{\pi}(Y - \mu_1)\right\}\right)$$

The logic for the $\widehat{\mu}_0$ part is analogous.

□

Theorem 2.1 is powerful in showing that, in Bernoulli trials, mean counterfactuals can be estimated very precisely (i.e., with zero bias and variance that scales like $1/n$) using no assumptions other than consistency and finite variance. In other words: randomization allows accurate and essentially assumption-free causal inference!

Similarly, Theorems 2.1 and 2.2 also pave the way for inference, in the form of confidence intervals and hypothesis tests. Namely, finite sample confidence intervals could be constructed based on Theorem 2.1 using bounds on the conditional variances σ_a^2, and Theorem 2.2 implies for example that an asymptotic 95% CI is given by

$$\widehat{\psi} \pm \left(\frac{1.96}{\sqrt{n}}\right) \widehat{\text{sd}} \left\{ \frac{A(Y - \widehat{\mu}_1)}{\pi} - \frac{(1-A)(Y - \widehat{\mu}_1)}{1-\pi} \right\}.$$

Remark 2.6. We saw above that the asymptotic variance of the difference-in-means estimator in a Bernoulli experiment is given by

$$\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}.$$

One interesting thing to note about this variance comes the perspective of experimental design: what is the best choice of π for optimizing efficiency? In fact, it is easy to show that

$$\arg\min_{\pi} \left(\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi} \right) = \frac{\sigma_1}{\sigma_0 + \sigma_1}$$

so for optimal efficiency the proportion treated should match the standard deviation of treated outcomes, as a fraction of the total standard deviation for treated and untreated outcomes. This is intuitive – if outcomes are more variable among treated patients than controls (i.e., $\sigma_1 > \sigma_0$) then more patients should be assigned to treatment to counterbalance the extra noise.

2.4.2 Sample versus Population Effects

Here we point out an interesting connection between sample effect estimation in completely randomized experiments and population effect estimation in Bernoulli experiments.

Based on Theorem 2.2, an asymptotic 95% CI for ψ in a Bernoulli experiment is given by

$$\widehat{\psi} \pm 1.96 \sqrt{\frac{\widehat{\sigma}_1^2}{n\widehat{\pi}} + \frac{\widehat{\sigma}_0^2}{n(1-\widehat{\pi})}}$$

where $\widehat{\sigma}_a^2 \equiv \sigma_n^2(y^a)$ is the usual sample variance among the treated ($a = 1$) and controls ($a = 0$), which we used in our analysis of the difference-in-means as an estimator of the *sample* average effect in completely randomized experiments (e.g., Proposition 2.3).

In fact, $\widehat{\psi}$ is the exact same point estimate of the sample effect that we analyzed in completely randomized experiments, and similarly the exact same confidence interval

$$\widehat{\psi} \pm 1.96 \sqrt{\frac{\widehat{\sigma}_1^2}{n\widehat{\pi}} + \frac{\widehat{\sigma}_0^2}{n(1-\widehat{\pi})}}$$

is also valid (possibly conservative) in completely randomized experiments, guaranteeing at least 95% coverage of the sample effect. (This results from using the naive bound of $\sigma_n^2(y^1 - y^0) \geq 0$ as in (2.2)).

Thus, not only is the estimator for the population effect exactly the same as that for the sample effect, but confidence intervals for the population effect are also valid for the sample effect, being at worst conservative. This is an archetypal example of how finite-sample and population-based frameworks can coincide.

Note that, although population-based confidence intervals are valid for sample effects, the converse is not necessarily true: it is easier to estimate sample effects, in the sense that the same estimators have smaller variances relative to sample versus population effects. Thus a confidence interval for a sample effect may not be valid for a population effect. For example, Imbens [2004] shows that

$$\mathbb{E}\{(\widehat{\psi} - \psi_n)^2\} = \mathbb{E}\{(\widehat{\psi} - \psi)^2\} - \frac{\text{var}(Y^1 - Y^0)}{n} + o(1/n)$$

so that the difference-in-means has smaller variance when estimating the sample effect ψ_n. For some intuition, imagine both potential outcomes were observed for each subject: then the sample effect would be estimated without error, but not the population effect.

2.4.3 Difference-in-Means versus Horvitz-Thompson

Note that the difference-in-means estimator is given by

$$\widehat{\psi} = \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)} = \mathbb{P}_n\left\{\left(\frac{A}{\widehat{\pi}}\right)Y - \left(\frac{1-A}{1-\widehat{\pi}}\right)Y\right\}$$

which suggests a different version, where we replace the estimated proportion treated $\widehat{\pi}$ with its known population value π:

$$\widehat{\psi}_{ht} = \mathbb{P}_n\left\{\left(\frac{A}{\pi}\right)Y - \left(\frac{1-A}{1-\pi}\right)Y\right\}$$

This estimator is known as the Horvitz-Thompson estimator, hence the *ht* subscript.

Since we are replacing an estimated quantity $\widehat{\pi}$ with its known value π, it seems as if we should gain efficiency. Is this actually true?

It is straightforward to check that the Horvitz-Thompson estimator is also unbiased and consistent; and since it is exactly equal to a sample average, we can apply the central limit theorem to immediately obtain

$$\sqrt{n}(\widehat{\psi}_{ht} - \psi) \rightsquigarrow N\left(0, \text{var}\left\{\left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)Y\right\}\right)$$

Now which estimator should we use: difference-in-means or Horvitz-Thompson? Both are unbiased, root-n consistent, and asymptotically normal. Is our intuition correct that it is beneficial to replace the estimate $\widehat{\pi}$ with its known value π? To answer this we will compare asymptotic variances.

Let $\phi = \frac{A}{\pi}(Y - \mu_1) - \frac{1-A}{1-\pi}(Y - \mu_0)$ and $\phi_{ht} = \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right) Y$ denote the functions whose variances correspond to the asymptotic variances of $\widehat{\psi}$ and $\widehat{\psi}_{ht}$. Then we have

$$\begin{aligned}\operatorname{var}(\phi_{ht}) &= \operatorname{var}\left(\phi + \frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right) \\ &= \operatorname{var}(\phi) + \operatorname{var}\left(\frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right)\end{aligned}$$

where the last line follows since $\mathbb{E}(\phi \mid A) = 0$ implies that

$$\operatorname{cov}\left(\phi, \frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right) = 0$$

by iterated expectation.

Therefore

$$\operatorname{var}(\phi_{ht}) \geq \operatorname{var}(\phi)$$

and thus the Horvitz-Thompson estimator is *less efficient* than the difference-in-means. This is counterintuitive: here replacing an estimated quantity with its known population counterpart actually reduces efficiency! Usually when we estimate things we get something *less precise* than if we just used the true quantity.

Unfortunately, I do not know of a very satisfying intuitive explanation of this paradox. One way to think about it is as follows: rather than viewing $\widehat{\psi}_{ht}$ as replacing an estimated quantity with a known quantity, one can instead view it as moving away from the sample average $\widehat{\psi} = \widehat{\mu}_1 - \widehat{\mu}_0$ with a noisier version

$$\widehat{\psi}_{ht} = \left(\frac{\widehat{\pi}}{\pi}\right)\widehat{\mu}_1 - \left(\frac{1-\widehat{\pi}}{1-\pi}\right)\widehat{\mu}_0$$

which should degrade performance, merely since sample averages are efficient estimators of means. In other words, the Horvitz-Thompson estimator is using the expected number of treated $n\pi$ rather than the actual number $n\widehat{\pi}$, so that when the actual number differs from its expectation, the averages are not correctly weighted.

Chapter 3

Randomized Experiments with Covariates

3.1 Identification with Covariates

So far we have considered settings where we have access to an iid sample

$$(A_1, Y_1), ..., (A_n, Y_n) \sim \mathbb{P}$$

but it is very common to also observe auxiliary covariate information (e.g., demographics like age or gender, or baseline outcome measures, etc.). Thus in practice we often have an iid sample

$$(X_1, A_1, Y_1), ..., (X_n, A_n, Y_n) \sim \mathbb{P}$$

for covariates or features $X \in \mathbb{R}^d$.

For now we will continue to pursue the experimental setting in which we can assume

1. consistency: $Y = AY^1 + (1 - A)Y^0$
2. randomization: $A \perp\!\!\!\perp (X, Y^a)$ for $a \in \{0, 1\}$ with $\mathbb{P}(A = 1 \mid X) = \pi$

Our goal is still to estimate the average treatment effect

$$\psi = \mathbb{E}(Y^1 - Y^0),$$

i.e., the mean outcome in the population if all versus none were treated.

The questions we consider here are: Does our identification result $\psi = \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0)$ without covariates still hold? Are there any new identification results that the covariates buy us? We will see that the answer to both questions is: yes.

We will require the following standard independence result.

Proposition 3.1. *If $U \perp\!\!\!\perp (V, W)$ then $U \perp\!\!\!\perp W$ and $U \perp\!\!\!\perp V \mid W$.*

Proof. Since $U \perp\!\!\!\perp (V, W)$ means $p(u, v, w) = p(u)p(v, w)$ we have

$$p(u, w) = \int p(u, v, w)\, dv = \int p(u)p(v, w)\, dv = p(u)p(w)$$

so that $U \perp\!\!\!\perp W$, showing the first part. For the second part note

$$p(u, v \mid w) = p(u \mid v, w)p(v \mid w) = p(u)p(v \mid w) = p(u \mid w)p(v \mid w).$$

□

Proposition 3.2. *Assume consistency and randomization as given above. Then*

$$\begin{aligned}\mathbb{E}(Y^1 - Y^0) &= \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0) \\ &= \mathbb{E}\Big\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\Big\}\end{aligned}$$

where

$$\mathbb{E}\Big\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\Big\} = \int \Big\{\mathbb{E}(Y \mid X = x, A = 1) - \mathbb{E}(Y \mid X = x, A = 0)\Big\}\, d\mathbb{P}(x).$$

Proof. Proposition 3.1 tells us that $A \perp\!\!\!\perp (X, Y^a) \implies A \perp\!\!\!\perp Y^a$ and $A \perp\!\!\!\perp Y^a \mid X$, and we already know $A \perp\!\!\!\perp Y^a$ implies

$$\mathbb{E}(Y^a) = \mathbb{E}(Y^a \mid A = a) = \mathbb{E}(Y \mid A = a)$$

by randomization and consistency. For the second identification result note

$$\mathbb{E}(Y^a) = \mathbb{E}\{\mathbb{E}(Y^a \mid X)\} = \mathbb{E}\{\mathbb{E}(Y^a \mid X, A = a)\} = \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}$$

where the first equality follows by iterated expectation, the second by $A \perp\!\!\!\perp Y^a \mid X$, and the third by consistency. □

The two identification results above suggest (at least) two different estimators for, for example, $\psi_1 = \mathbb{E}(Y^1)$, namely:

$$\widehat{\psi}_1 = \mathbb{P}_n(Y \mid A = 1) \;\text{ versus }\; \widehat{\psi}_1 = \mathbb{P}_n\{\widehat{\mathbb{E}}(Y \mid X, A = 1)\}$$

The questions we pursue here are: Which estimator is "better"? Should we incorporate the covariate information? How?

3.2 Logistic Regression & Collapsibility

For the time being, suppose $Y \in \{0, 1\}$ is a binary outcome. Maybe the most common approach in practice is to assume the logistic regression model

$$\text{logit}\ \mathbb{P}(Y = 1 \mid X, A) = \beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X$$

and call β_1 "the effect" of treatment. What "effect" does this actually represent?

First: this is likely not an effect at all, because *the model is probably wrong*. For better or worse, nature does not care about our logistic regressions. We fit logistic regression models because they are fast and easy, not because they are realistic. In reality any given model probably leaves out important covariate interactions, higher-order terms, covariate-treatment interactions, non-logit links, etc.

In other words, it is pretty presumptuous to assume we know the *exact* functional form explaining how covariates relate to the outcome, up to some finite-dimensional parameter. Nature is probably too complex for that. And thus if the model is wrong, β_1 is hard to interpret and not so meaningful – a projection at best.

Nevertheless, for the sake of argument assume that the logistic model is correct. Then

$$\exp(\beta_1) = \frac{\text{odds}(Y = 1 \mid X, A = 1)}{\text{odds}(Y = 1 \mid X, A = 0)} = \frac{\text{odds}(Y^1 = 1 \mid X)}{\text{odds}(Y^0 = 1 \mid X)}$$

where in the second equality we used consistency and randomization.

This is a *conditional odds ratio* (OR). Importantly

$$\mathbb{E}(Y^1 - Y^0) = \mathbb{P}(Y^1 = 1) - \mathbb{P}(Y^0 = 1) \neq \frac{\text{odds}(Y^1 = 1 \mid X)}{\text{odds}(Y^0 = 1 \mid X)}$$

so it is certainly not an average treatment effect (risk difference). In fact, *even if the model is correct*

$$\frac{\text{odds}(Y^1 = 1)}{\text{odds}(Y^0 = 1)} \neq \frac{\text{odds}(Y^1 = 1 \mid X)}{\text{odds}(Y^0 = 1 \mid X)}$$

so it is not even a population odds ratio effect (even if the conditional OR is constant!). This follows since

$$\begin{aligned}
\text{odds}(Y^1 = 1) &= \frac{\mathbb{P}(Y^1 = 1)}{\mathbb{P}(Y^1 = 0)} = \frac{\mathbb{P}(Y = 1 \mid A = 1)}{\mathbb{P}(Y = 0 \mid A = 1)} \\
&= \frac{\mathbb{E}\{\mathbb{P}(Y = 1 \mid X, A = 1)\}}{\mathbb{E}\{\mathbb{P}(Y = 0 \mid X, A = 1)\}} = \frac{\mathbb{E}\{\text{expit}(\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} X)\}}{1 - \mathbb{E}\{\text{expit}(\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} X)\}} \\
&\neq \frac{\text{expit}\{\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} \mathbb{E}(X)\}}{1 - \text{expit}\{\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} \mathbb{E}(X)\}} = \exp\{\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} \mathbb{E}(X)\}
\end{aligned}$$

since $\mathbb{E}\{f(X)\} \neq f\{\mathbb{E}(X)\}$ for nonlinear f. This is called the problem of *non-collapsibility* [Freedman, 2008, Greenland et al., 1999]. We say the odds ratio is not collapsible since the average of the conditional ORs is not generally equal to the marginal OR.

In fact the marginal OR can be bigger/small than all of the conditional ORs. This is counterintuitive! Consider an example with half men and half women:

- men have 20% chance of heart attack when treated, 50% when not

- women have 2% chance of heart attack when treated, 8% when not

What are the ORs in this setup?

$$OR(HA \mid men) = \frac{\text{odds}(HA^1 \mid men)}{\text{odds}(HA^0 \mid men)} = \frac{0.2/0.8}{0.5/0.5} = 0.25$$
$$OR(HA \mid women) = \frac{\text{odds}(HA^1 \mid women)}{\text{odds}(HA^0 \mid women)} = \frac{.02/0.98}{0.08/0.92} = 0.235$$
$$OR(HA) = \frac{\text{odds}(HA^1)}{\text{odds}(HA^0)} = \frac{0.11/0.89}{0.29/0.71} = 0.303$$

Therefore the marginal OR is larger than that in either stratum!

The main take-away is that coefficients in general non-linear models are conditional and do not correspond to marginal (entire-population) effects (even if the model is correct, which is probably unlikely). This subtlety is often missed.

For example, suppose we did a study and obtained the above data. If we fit a (misspecified) logistic regression assuming the ORs were constant for men and women, we would report an OR of $\sim$0.24, but this overstates the OR we'd see if we gave entire population treatment versus not ($\sim$0.3). This problem can be exacerbated the more covariates there are, or of course if the model is misspecified.

However, this problem does not arise in a (correctly specified) linear model, e.g.,

$$\mathbb{E}(Y \mid X, A) = \beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X$$

If the model is correct, then

$$\beta_1 = \mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)$$

so the coefficient is a conditional effect.

However, under the linear model assumption and $A \perp\!\!\!\perp Y^a \mid X$

$$\mathbb{E}(Y^1 - Y^0) = \mathbb{E}\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\} = \beta_1$$

so the parameter is also a marginal effect.

We have seen that going after coefficients in nonlinear regression models can be sub-optimal in experiments. Namely, we typically have to assume the model is correct (a huge assumption) and, even if the model is correct, the coefficient in that case will be a conditional effect which does not correspond to a well-defined effect in the whole population.

In what follows we will discuss how to deal with the second issue, and then after that the first issue.

3.3 Recovering Population Effects via Regression

In the previous section we saw that, under parametric model assumptions (with randomization), the coefficient from a logistic regression model recovers a conditional odds ratio. Here we consider the question of how we might use this fit to estimate the marginal average treatment effect.

First a few basics. When we fit a logistic (or any other) regression model we are estimating the conditional expectation function

$$\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a).$$

For example in logistic regression we estimate the function μ with

$$\widehat{\mu}_a(x) = \text{expit}(\widehat{\beta}_0 + \widehat{\beta}_1 a + \widehat{\beta}_2^{\text{T}} x).$$

In R this function can be evaluated at the observed $(X_1, A_1), ..., (X_n, A_n)$ values with the `predict` command, as in:

```
lrmod <- glm(y ~ x+a, family=binomial)
muhat <- predict(lrmod, type="response")
```

Now recall that under the randomization assumption $A \perp\!\!\!\perp (Y^a, X)$ (together with consistency) the average treatment effect is given by

$$\psi = \mathbb{E}(Y^1 - Y^0) = \mathbb{E}\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\}$$

which suggests the estimator

$$\begin{aligned}\widehat{\psi} &= \mathbb{P}_n\Big\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\Big\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\Big\}\end{aligned}$$

This estimator is sometimes called the *plug-in*, *g-computation*, or *standardization* estimator. Conceptually, it is taking the estimated conditional effect $\widehat{\mu}_1(x) - \widehat{\mu}_0(x)$ and standardizing it to the (empirical) population distribution of covariates. You can also think of it as "imputing" an estimate of the effect $\mu_1(x) - \mu_0(x)$ for each person and averaging.

How would you compute this estimator in practice? First you can obtain predicted values under $A = 1$ and $A = 0$ separately, for everyone (regardless of actual observed treatment), then take the difference for each person, and average across people. Note that in R, the `predict` function outputs predicted values under the *observed* treatment; in contrast here one needs predicted values under $A = 1$ and $A = 0$ separately, so you need the `newdata` argument. Here is some example code for a simulated dataset:

```
> cbind(x,a,y)
                x a y
 [1,] -0.44577826 0 0
 [2,] -1.20585657 0 0
 [3,]  0.04112631 1 1
 [4,]  0.63938841 0 0
 [5,] -0.78655436 0 1
 [6,] -0.38548930 0 1
 [7,] -0.47586788 1 0
 [8,]  0.71975069 1 1
 [9,] -0.01850562 1 1
...
[100,]  2.01893816 0 1
>
> mumod <- glm(y~x+a, family=binomial)
>
> mu1hat <- predict(mumod, newdata=data.frame(x,a=1) ,type="response")
> mu0hat <- predict(mumod, newdata=data.frame(x,a=0), type="response")
>
> cbind(x,a,y, mu1hat, mu0hat, mu1hat-mu0hat)
              x a y    mu1hat    mu0hat
1   -0.44577826 0 0 0.8178912 0.5709608 0.2469305
2   -1.20585657 0 0 0.7793358 0.5113598 0.2679760
3    0.04112631 1 1 0.8397107 0.6081932 0.2315174
4    0.63938841 0 0 0.8635670 0.6522366 0.2113304
5   -0.78655436 0 1 0.8012901 0.5443888 0.2569013
6   -0.38548930 0 1 0.8207133 0.5756239 0.2450893
7   -0.47586788 1 0 0.8164699 0.5686286 0.2478412
8    0.71975069 1 1 0.8665332 0.6579775 0.2085556
9   -0.01850562 1 1 0.8371566 0.6036912 0.2334654
...
100  2.01893816 0 1 0.9073274 0.7436597 0.1636677
>
> mean(mu1hat-mu0hat)
[1] 0.2303284
```

So for the above simulated dataset, the estimated average treatment effect using logistic regression is $\widehat{\psi} = 0.23$.

Note there is no particular reason to favor logistic regression for constructing the regression estimates $\widehat{\mu}_a(x)$; one might instead consider linear regression, probit regression, regression trees, kernel estimators, splines, generalized additive models, the lasso, boosting, random forests, neural networks, deep learning, etc.

3.3.1 Properties of the Plug-in Estimator

In this section we analyze the simple plug-in estimator $\widehat{\psi}$. We do so by finding answers to three standard questions: Is the estimator consistent? What is its convergence rate? What is its asymptotic limiting distribution?

Notice that if we let $\widehat{f} = \widehat{\mu}_1 - \widehat{\mu}_0$, then $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ and $\psi = \mathbb{E}(f)$. Thus our estimator is a sample average of an estimated function, and our target estimand is an expectation of the true function f. Thus its performance will be very closely tied to the errors in estimating the function f with $\widehat{f}$.

At this point it will be useful to take a slight detour and discuss properties of estimated functions, and introduce some new notation.

Estimated functions

First, since our analysis involves differences between the estimated function $\widehat{f}$ and the truth f, i.e., errors in estimating f with $\widehat{f}$, we will need a notion of consistency for random functions $\widehat{f}$. Recall in Chapter 1 we learned that a scalar (or Euclidean) estimator $\widehat{\psi}$ is consistent if $\widehat{\psi} - \psi = o_{\mathbb{P}}(1)$, i.e., if $\widehat{\psi}$ converges to ψ in probability.

For functions we can define an appropriate (scalar) distance measure, and then consistency will be defined as in the scalar case. Some popular distance measures for functions are:

- L_1 distance: $\|\widehat{f} - f\|_1 = \int |\widehat{f}(x) - f(x)| \, d\mathbb{P}(x)$
- L_2 distance: $\|\widehat{f} - f\|_2 = \sqrt{\int \{\widehat{f}(x) - f(x)\}^2 \, d\mathbb{P}(x)}$
- L_∞ distance: $\|\widehat{f} - f\|_\infty = \sup_{x \in \mathcal{X}} |\widehat{f}(x) - f(x)|$

Note that all of these distances are themselves random variables, since they depend on the estimated $\widehat{f}$. Now we are ready to define consistency of an estimated function.

Definition 3.1. An estimated function $\widehat{f}(x)$ is consistent for a fixed target $f(x)$ in distance measure $d(\cdot, \cdot)$ if

$$d(\widehat{f}, f) = o_{\mathbb{P}}(1).$$

Similarly, $\widehat{f}$ converges at rate $r_n \to \infty$ to f in distance d if

$$d(\widehat{f}, f) = o_{\mathbb{P}}(1/r_n).$$

In addition to having a notion of consistency or convergence for estimated functions $\widehat{f}$, it will also be useful for us to have some special notation for the expected value over a random function's argument, conditioning on the randomness in the function.

Definition 3.2. For an estimated function $\widehat{f}(x)$ built from a sample $Z^n = (Z_1, ..., Z_n)$ we use the notation

$$\mathbb{P}(\widehat{f}) = \mathbb{P}\{\widehat{f}(Z)\} \equiv \int \widehat{f}(z)\ d\mathbb{P}(z) = \mathbb{E}\left\{\widehat{f}(Z) \;\middle|\; Z^n\right\}$$

to denote expectations over a new independent observation Z, conditioning on the sample Z^n.

Remark 3.1. The heuristic interpretation of $\mathbb{P}(\widehat{f})$ is as follows: you construct the function $\widehat{f}(z)$ from a sample Z^n, and then take its average over new repeated independent draws of the argument Z. It is important to note that, for a *fixed* function $f(z)$ we have

$$\mathbb{E}\{f(Z)\} = \mathbb{P}(f)$$

whereas for a random estimated function $\widehat{f}(z)$ depending on a sample Z^n, we have

$$\mathbb{E}\{\widehat{f}(Z)\} = \mathbb{E}\left[\mathbb{E}\left\{\widehat{f}(Z) \;\middle|\; Z^n\right\}\right] \neq \mathbb{P}(\widehat{f}).$$

In particular, the quantity $\mathbb{P}(\widehat{f})$ on the right-hand-side is random (through its dependence on $\widehat{f}$ and Z^n), whereas the quantities on the left-hand-side are fixed.

Back to the standardization estimator

Now we are ready to proceed with investigating the plug-in estimator $\widehat{\psi} = \mathbb{P}_n(\widehat{\mu}_1 - \widehat{\mu}_0)$ of the average treatment effect: in particular the three fundamental properties of consistency, rate of convergence, and limiting distribution.

The first tells us whether the estimator is at least converging to the correct target as sample size increases (the lowest bar we would hope an estimator would clear), the second how quickly this convergence occurs (i.e., how much information in the sample the estimator makes use of), and the third whether the estimator is well-behaved enough to give us hope for constructing confidence intervals and doing inference.

At a high level, our goal is to write $\widehat{\psi} - \psi$ as a (centered) sample average, plus some noise. We know how to analyze sample averages, since for any fixed function g of the iid observations Z, we have $(\mathbb{P}_n - \mathbb{P})g(Z) = (\mathbb{P}_n - \mathbb{E})g(Z) = O_{\mathbb{P}}(1/\sqrt{n})$ and in particular

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P})g(Z) \rightsquigarrow N\Big(0, \mathrm{var}\{g(Z)\}\Big)$$

by the central limit theorem. Therefore the problem will be reduced to analyzing whatever the noise is.

First we will introduce a foundational decomposition for $\widehat{\psi}$ (in fact, for any estimator that takes a similar form), which will be crucial for many estimators we analyze throughout the course.

Lemma 3.1. *Let $\widehat{\psi} = \mathbb{P}_n(\widehat{f}) = \frac{1}{n}\sum_i \widehat{f}(Z_i)$ be an estimator of the generic expectation $\psi = \mathbb{P}(f) = \mathbb{E}\{f(Z)\}$ based on n samples $(Z_1, ..., Z_n)$, where $\widehat{f}$ can be any estimator and $f : \mathcal{Z} \mapsto \mathbb{R}$ any function. Then we have the decomposition*

$$\widehat{\psi} - \psi = Z^* + T_1 + T_2 \tag{3.1}$$

where

$$\begin{aligned} Z^* &= (\mathbb{P}_n - \mathbb{P})f \\ T_1 &= (\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) \\ T_2 &= \mathbb{P}(\widehat{f} - f). \end{aligned}$$

Proof. We have

$$\begin{aligned} \widehat{\psi} - \psi &= \mathbb{P}_n(\widehat{f}) - \mathbb{P}(f) \\ &= (\mathbb{P}_n - \mathbb{P})\widehat{f} + \mathbb{P}(\widehat{f} - f) \\ &= (\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) + (\mathbb{P}_n - \mathbb{P})f + \mathbb{P}(\widehat{f} - f) \\ &\equiv T_1 + Z^* + T_2 \end{aligned}$$

where the first line follows by definition, the second by adding and subtracting $\mathbb{P}(\widehat{f})$ (which we recall is not the same as $\mathbb{E}(\widehat{f})$), and the third by adding and subtracting the quantity $(\mathbb{P}_n - \mathbb{P})f = (\mathbb{P}_n - \mathbb{E})f = \frac{1}{n}\sum_i [f(X_i) - \mathbb{E}\{f(X_i)\}]$. □

Remark 3.2. Lemma 3.1 immediately applies to the plug-in estimator $\widehat{\psi}$ if as before we let $\widehat{f} = \widehat{\mu}_1 - \widehat{\mu}_0$ so that the estimator $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ is a sample average of the estimated function $\widehat{f}$, and the target parameter $\psi = \mathbb{E}(f)$ is the population expectation of the true function f.

Lemma 3.1 has achieved our goal of writing our estimator as a centered sample average plus noise. The first term Z^* in (3.1) is a nice centered sample average, and so by the central limit theorem it behaves as a normally distributed variable with variance $\mathrm{var}(f)/n$, up to error $o_{\mathbb{P}}(1/\sqrt{n})$. Thus our problem is reduced to analyzing the two noise terms, denoted T_1 and T_2.

Remark 3.3. Note that the decomposition in (3.1) only relied on the estimator being a sample average of an estimated function $\widehat{f}$, and on the estimand being an expectation of a true function f. There was nothing special about ψ being the average treatment effect, or f being a regression function. We will see the decomposition (3.1) repeatedly throughout the course, since many estimands are expected values of (sometimes complicated) generic functions, which can be estimated by corresponding sample averages of estimates of these functions.

Now we will analyze the noise terms T_1 and T_2.

It turns out that the term T_1 is typically of smaller order than even the Z^* term. In fact, $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$ under some weak regularity conditions, as long as $\widehat{f}$ is consistent for f in L_2 norm, i.e., as long as

$$\|\widehat{f} - f\|_2^2 = \int \{\widehat{f}(x) - f(x)\}^2 \, d\mathbb{P}(x) = o_{\mathbb{P}}(1).$$

We will prove this rigorously later in the course (for a sneak peek see Kennedy et al. [2019a]). Intuitively, however, this should not be too surprising, since T_1 is a centered sample average (just like Z^*), but in fact the quantity it is averaging is shrinking to zero with n (as long as $\widehat{f}$ is tending to f). This is like taking larger and larger centered sample averages of a random variable whose variance shrinks with n.

Now we turn to the last noise term T_2, which is the really interesting one. For many estimators we discuss in the course, the T_2 term will be particularly crucial, driving the rate of convergence and limiting distribution.

3.3.2 The Parametric Plug-in Estimator

First we consider analyzing $T_2 = \mathbb{P}(\widehat{f} - f)$ in the case where $\widehat{f}$ is estimated with a (correct) parametric model, i.e., where

$$\widehat{f}(x) = f(x; \widehat{\beta})$$

for some finite-dimensional parameter $\beta \in \mathbb{R}^p$. For example, when using the logistic regression model as before, we would have $f(x;\beta) = \text{expit}(\beta_0+\beta_1+\beta_2^{\text{T}}x)-\text{expit}(\beta_0+\beta_2^{\text{T}}x)$.

Note in the parametric case we can view

$$\begin{aligned} T_2 &= \mathbb{P}(\widehat{f} - f) \\ &= \int \{f(x;\widehat{\beta}) - f(x;\beta)\} \, d\mathbb{P}(x) \\ &\equiv g(\widehat{\beta}) - g(\beta) \end{aligned}$$

as a simple difference in functions $\widehat{\beta}$ and β, where the function g will be smooth if f is. Therefore we will first understand the error between $\widehat{\beta}$ and β, and then use the delta method.

For most smooth parametric models, the estimator $\widehat{\beta}$ will solve an estimating equation based on some mean-zero estimating function m that is smooth in β. For example, the logistic regression estimator solves an estimating equation based on the estimating function (or score function)

$$m(Z;\beta) = \begin{pmatrix} 1 \\ A \\ X \end{pmatrix} \Big\{ Y - \text{expit}(\beta_0 + \beta_1 A + \beta_2^{\text{T}} X) \Big\}.$$

so that

$$\mathbb{P}_n\{m(Z;\widehat{\beta})\} = 0$$

by definition. The next standard result shows that such solutions to finite-dimensional estimating equations behave like sample averages.

Lemma 3.2. *Suppose the estimator $\widehat{\beta} \in \mathbb{R}^p$ solves an estimating equation so that*

$$\mathbb{P}_n\{m(Z;\widehat{\beta})\} = 0.$$

Assume $m(z;\beta) \in \mathbb{R}^p$ is Lipschitz in β, and that $\mathbb{E}\{m(z;\beta)\}$ is differentiable at the population β satisfying $\mathbb{E}\{m(Z;\beta)\} = 0$ with nonsingular derivative matrix. Then

$$\widehat{\beta} - \beta = (\mathbb{P}_n - \mathbb{P}) \left[\left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z;\beta) \right] + o_{\mathbb{P}}(1/\sqrt{n}), \tag{3.2}$$

Proof. See Theorem 5.23 of van der Vaart [2000]. □

Corollary 3.1. *Under the conditions of Lemma 3.2, the estimating equation estimator $\widehat{\beta}$ is root-n consistent and asymptotically normal.*

Now we have all the tools we need to analyze the quantity $\mathbb{P}(\widehat{f} - f)$ and thus the estimator $\widehat{\psi}$ in the parametric case.

Theorem 3.1. *Let $f(x) = \mu_1(x) - \mu_0(x)$ and $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$, so that $\psi = \mathbb{E}\{f(X)\}$ is the average treatment effect. Assume the parametric model*

$$\mu_a(x) = \mu_a(x;\beta)$$

for some $\beta \in \mathbb{R}^p$, and that the estimator $\widehat{\beta}$ satisfies the conditions of Lemma 3.2. Then

$$\widehat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P})g(Z;\beta) + o_{\mathbb{P}}(1/\sqrt{n})$$

where

$$g(z;\beta) = f(x;\beta) + \frac{\partial \mathbb{E}\{f(X;\beta)\}}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(z;\beta)$$

and so is root-n consistent and asymptotically normal.

Proof. By Lemma 3.1 we have

$$\widehat{\psi} - \psi = Z^* + T_1 + T_2$$

where $Z^* = (\mathbb{P}_n - \mathbb{P})f$ and T_1 and T_2 defined accordingly. By Lemma 3.2 we have

$$\widehat{\beta} - \beta = (\mathbb{P}_n - \mathbb{P}) \left[\left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z;\beta) \right] + o_{\mathbb{P}}(1/\sqrt{n})$$

which also is enough to imply $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$. Further by the delta method we have

$$\begin{aligned} T_2 &= g(\widehat{\beta}) - g(\beta) \\ &= (\mathbb{P}_n - \mathbb{P}) \left[\frac{\partial g(\beta)}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z;\beta) \right] + o_{\mathbb{P}}(1/\sqrt{n}) \end{aligned}$$

for $g(\beta) = \mathbb{E}\{f(Z;\beta)\}$. Combining the terms gives the result. □

To summarize, when $\widehat{\mu}$ is estimated with a correct parametric model, the resulting plug-in estimator $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ for $f = \mu_1 - \mu_0$ is root-n consistent for the causal effect ψ and asymptotically normal.

When the parametric model for μ is correct, this plug-in estimator is most efficient (this follows from classical low-dimensional parametric maximum likelihood theory); the intuition is that with a correct simple model we can "predict" the treatment effect much more precisely than say with the difference-in-means estimator. Further confidence intervals can be constructed using estimates of the closed-form asymptotic variance given above, or via bootstrap (which is typically easier).

Of course, these days we rarely believe our parametric models are actually correct, especially when X contains some continuous covariates or is high-dimensional. At best such models may be a modestly biased approximation, but at worst, when very misspecified they may turn our estimation procedure into garbage, yielding estimates that are not only far away from the truth, but to an unknown extent.

3.3.3 The Nonparametric Plug-in

This begs the question of how the plug-in estimator would behave if we used a more flexible estimator to construct $\widehat{\mu}$, say random forests or the lasso or deep learning.

In this case, of course the central limit theorem term Z^* in our decomposition (3.1) is still going to behave as a mean-zero normally distributed random variable with variance $\mathrm{var}(f)/n$, since it does not depend on the estimated $\widehat{f}$. Further, even when μ is treated as a potentially infinite-dimensional function and estimated flexibly and data-adaptively, the term T_1 can still be of smaller order (though we may need to use sample splitting, as will be discussed in detail later in the course).

Unfortunately the picture is nowhere near as rosy for the important T_2 term in (3.1). If all we know about the flexible estimator $\widehat{f}$ are high-level rates of convergence, say in L_2 norm, then all we can say about T_2 is

$$T_2 = \mathbb{P}(\widehat{f} - f) \leq \sqrt{\mathbb{P}\{(\widehat{f} - f)^2\}} = \|\widehat{f} - f\|_2$$

where the second inequality uses Cauchy-Schwarz. This means in general we would expect the plug-in estimator $\widehat{\psi}$ to *inherit* the (typically slow) rate of convergence of the nonparametric estimator $\widehat{f}$.

This is a problem since for most realistic infinite-dimensional function classes the L_2 norm will be far away from $1/\sqrt{n}$. For example when f lies in a Hölder class with index s (i.e., all partial derivatives up to order s exist and s^{th} derivatives Lipschitz) then for *any* estimator $\widehat{f}$ the rate cannot be any faster than

$$\|\widehat{f} - f\|_2 \gtrsim n^{-s/(2s+d)}$$

uniformly over the Hölder class [Tsybakov, 2009]; note this rate is always slower than $\sqrt{n}$. For example, suppose we are only willing to assume our regression functions have $s = 2$ derivatives, and we have $d = 16$ covariates. Then the best achievable rate is $n^{-1/10}$. Neural network classes are known for yielding dimension-independent rates [Györfi et al., 2002], but even these are $n^{-1/4}$, somewhat of a far cry from $1/\sqrt{n}$.

Further, when $\widehat{\mu}$ is estimated flexibly with modern nonparametric tools, we do not only pay a price in the rate of convergence – it will generally also be true that, even if we can derive a tractable limiting distribution, there will be some smoothing bias, so confidence intervals will not be correctly centered (even using the bootstrap) and thus will not cover at the nominal level. However, often complex nonparametric estimators do not even yield tractable limiting distributions, even uncentered.

3.4 Efficient Model-Free Estimation

At this point we find ourselves in a bit of a quandary. We could use the simple difference-in-means estimator, which is root-n consistent and asymptotically normal *under no modeling assumptions*; however it completely ignores covariate information and so may be quite inefficient relative to other estimators. Alternatively we could model the regression function and use the plug-in estimator. However if we use parametric models to achieve root-n rates and small confidence intervals, we are putting ourselves at great risk of bias due to model misspecification; on the other hand, if we model the regression functions nonparametrically, letting the data speak for themselves, then we will typically suffer from the curse of dimensionality and be subject to slow rates of convergence, and at a loss for confidence intervals and inference.

What should we do? Is there any way to get the best of both worlds, using the covariates to gain efficiency over the difference-in-means estimator, but retaining its model-free benefits and not risking bias?

3.4.1 The Doubly Robust Estimator

It turns out there exists a bias-corrected estimator, whose validity is based on randomization, yet which can incorporate regression predictions to increase efficiency:

$$\widehat{\psi} = \mathbb{P}_n \left[\left\{ \widehat{\mu}_1(X) - \widehat{\mu}_0(X) \right\} + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi} \right) \left\{ Y - \widehat{\mu}_A(X) \right\} \right] \tag{3.3}$$

where $\widehat{\mu}_a(x)$ is an estimate of the regression function $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$ and $\pi = \mathbb{P}(A = 1)$ is the (known) randomization probability.

The estimator (3.3) can be viewed as the plug-in estimator $\mathbb{P}_n(\widehat{\mu}_1 - \widehat{\mu}_0)$ plus a "correction" term that incorporates the randomization probabilities π. It goes by various names, including:

- model-assisted Horvitz-Thompson,
- bias-corrected plug-in,
- semiparametric or semiparametric efficient,
- augmented inverse-probability-weighted (AIPW),
- doubly robust.

We will see variants of the estimator (3.3) throughout the course, and will mostly refer to it as doubly robust. It has an interesting and somewhat difficult-to-trace history across subfields of statistics. Here is an abbreviated and limited portion of its path across the literature:

- In survey sampling problems, Cochran [1977] and others used simple regression models in an agnostic way to improve the efficiency of the unbiased Horvitz-Thompson estimator from 1952.
- Robins and Rotnitzky [1995], Robins et al. [1994, 1995] studied efficient semiparametric estimation in general missing data problems (extending work by Bickel et al. [1993] and Pfanzagl [1982] and others), and presented a version of this estimator (3.3) where nuisance quantities were estimated with parametric models.
- Robins and Wang [2000] started referring to the estimator (3.3) as "doubly protected", and Robins and Rotnitzky [2001] and Bang and Robins [2005] as "doubly robust".
- In a series of papers, Tsiatis and colleagues [Davidian et al., 2005, Leon et al., 2003, Yang and Tsiatis, 2001, Zhang et al., 2008] applied the theory from Robins and others to randomized experiments, focusing on efficiency concerns. These papers are a nice introduction to the estimator (3.3) in the experimental setup.
- In the early to mid 2000s, van der Laan and Robins [2003] and others started developing theory for the case where nuisance estimators such as $\widehat{\mu}_a$ are estimated nonparametrically.

- The estimator and related methods have been recently re-discovered in the econometrics world [Chernozhukov et al., 2018], with more of a focus on high-dimensional sparse models.

In fact it can be shown that any (regular) $\sqrt{n}$-consistent and asymptotically normal estimator can be written in the form (3.3), for some choice of $\widehat{\mu}_a$. So in fact we have already seen some variants of it, e.g.:

- The difference-in-means estimator is recovered if $\widehat{\mu}_a = \mathbb{P}_n(Y \mid A = a)$, and
- the Horvitz-Thompson or inverse-probability-weighted estimator if $\widehat{\mu}_a = 0$.

In fact, shortly we will study some cases where, surprisingly, the parametric plug-in takes this form with for example $\widehat{\mu}_a = g(\widehat{\beta}_0 + \widehat{\beta}_1 a + \widehat{\beta}_2^{\mathrm{T}} x)$. This is one of the reasons it is a bit unclear where the estimator originated, since it includes many variants as a special case.

Remark 3.4. As we did above, at several points in this section we will refer to *regular* estimators. A more detailed discussion will come later, but for the time being a regular estimator can be taken to mean an estimator whose limiting distribution is insensitive to local perturbations of the data-generating process. Imposing regularity rules out *super-efficient* estimators, for example, which trade very good performance at a particular $\mathbb{P}$ for very bad performance "near" $\mathbb{P}$. More discussion can be found in Tsiatis [2006] and van der Vaart [2000].

As mentioned earlier, the estimator (3.3) can be interpreted as a corrected version of the plug-in estimator $\widehat{\psi}_{pi} = \mathbb{P}_n(\widehat{\mu}_1 - \widehat{\mu}_0)$ since

$$\widehat{\psi} = \widehat{\psi}_{pi} + \mathbb{P}_n \left[\left(\frac{A}{\pi} - \frac{1-A}{1-\pi} \right) \left\{ Y - \widehat{\mu}_A(X) \right\} \right].$$

We will see how the correction term removes any bias afflicting the regression estimator $\widehat{\mu}_a$. The doubly robust estimator can also be viewed as a corrected (of "augmented") version of the inverse-probability weighted (Horvitz-Thompson) estimator $\widehat{\psi}_{ipw} = \mathbb{P}_n\{\left(\frac{AY}{\pi} - \frac{1-A}{1-\pi}\right) Y\}$ since

$$\widehat{\psi} = \widehat{\psi}_{ipw} + \mathbb{P}_n \left[\left(1 - \frac{A}{\pi} \right) \widehat{\mu}_1(X) - \left(1 - \frac{1-A}{1-\pi} \right) \widehat{\mu}_0(X) \right].$$

We know from the previous chapter that $\widehat{\psi}_{ipw}$ is already unbiased; thus the above augmentation term is reducing variance rather than bias.

Here is example code showing how to correct the plug-in estimator we constructed earlier:

```
> cbind(x,a,y)[1:5,]
                x a y
 [1,] -0.44577826 0 0
 [2,] -1.20585657 0 0
 [3,]  0.04112631 1 1
 [4,]  0.63938841 0 0
 [5,] -0.78655436 0 1
>
> mumod <- glm(y~x+a, family=binomial)
> mu1hat <- predict(mumod, newdata=data.frame(x,a=1) ,type="response")
> mu0hat <- predict(mumod, newdata=data.frame(x,a=0), type="response")
>
> pi <- 0.5; muahat <- a*mu1hat + (1-a)*mu0hat
>
> mean( (mu1hat-mu0hat) + (a/pi - (1-a)/(1-pi)) * (y-muahat) )
[1] 0.2303284
```

Remark 3.5. Note that the doubly robust estimator requires no extra model fitting beyond that already required to construct the plug-in estimator.

A natural question about the doubly robust estimator is: where does the correction come from, and why does it take that specific form? A complete answer to this is highly non-trivial; we will pursue it in depth in later chapters. However some short discussion is still useful. The form of the correction comes from nonparametric efficiency theory for functional estimation [Bickel et al., 1993, Tsiatis, 2006, van der Laan and Robins, 2003], and there are two high-level heuristics for thinking about it. The first is that the average treatment effect parameter $\psi = \psi(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}(\mu_1 - \mu_0)$ is a "smooth" functional, when viewed as a map from probability distributions $\mathbb{P}$ to the real line; and this smoothness allows for convenient and effective bias correction. The second is that a randomized experiment with known treatment mechanism leads to a semiparametric model for the distribution $\mathbb{P}$ from which we sample: part of the distribution $\mathbb{P}$ is known (the conditional distribution of treatment given any covariates) while the rest is left unrestricted (the covariate distribution and the conditional distribution of the outcome given covariates and treatment). Under this semiparametric model, one can use tools from efficiency theory to derive the form of *all* possible (regular) asymptotically normal estimators of the parameter ψ, and subsequently find the one with the smallest variance.

Answering the question of *why* the correction works is easier than answering where it comes from. This is the focus of the next section.

3.4.2 Properties of the Doubly Robust Estimator

Here we study the bias, variance, and limiting distribution of the estimator (3.3).

Remark 3.6. In this section we are going to consider the case where the regression estimator $\widehat{\mu}_a$ is constructed from a separate training sample D^n independent of the experimental sample $Z^n = \{(X_1, A_1, Y_1), ..., (X_n, A_n, Y_n)\}$. This setup can be accomplished easily in practice by simply randomly splitting the sample, and using half as D^n for training and the other half as Z^n for estimation. Note that in this case, variance results should really be framed in terms of $n/2$ instead of n; if this loss of efficiency is concerning to you, luckily there is an easy fix: after constructing the sample-split estimator, swap the samples, using Z^n for training and D^n for estimation, and then average the resulting estimators. This approach will recover full sample size efficiency.

There are two reasons for doing sample splitting: the first is that the analysis is more straightforward, and the second more important reason is that it prevents overfitting and allows for the use of arbitrarily complex estimators $\widehat{\mu}_a$ (e.g., random forests, boosting, neural nets). Without sample splitting, one would have to restrict the complexity of the estimator $\widehat{\mu}_a$ via empirical process conditions (e.g., via Donsker class or entropy restrictions). Intuitively, this is because the estimator $\widehat{\psi}$ is using the data twice: once to estimate the unknown function μ_a and once to estimate the bias correction. Sample splitting ensures that these tasks are accomplished independently.

As in our analysis of the plug-in estimator in the previous section, we note that our estimator can be written as a sample average of an estimated function. Namely $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ where now $\widehat{f} = f(\widehat{\mu}) \equiv f_1(\widehat{\mu}) - f_1(\widehat{\mu})$ for

$$f_a(\overline{\mu}) \equiv \overline{\mu}_a(X) + \frac{\mathbb{1}(A=a)}{\mathbb{P}(A=a)}\Big\{Y - \overline{\mu}_A(X)\Big\} \tag{3.4}$$

First we tackle the bias of $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$, *under no modeling assumptions.*

Theorem 3.2. *Consider an iid Bernoulli experiment with $\mathbb{P}(A=1) = \pi$. Then the doubly robust estimator $\widehat{\psi}$ in (3.3) is unbiased for the average treatment effect when the regression estimates $\widehat{\mu}_a$ are constructed from a separate independent sample.*

Proof. We will derive the bias for $\psi_1 = \mathbb{E}(Y^1)$ with $\widehat{\psi}_1 = \mathbb{P}_n(\widehat{f}_1)$ since the logic is exactly the same for $\mathbb{E}(Y^0)$ and the difference $\psi = \psi_1 - \psi_0$. First note that for any $\overline{\mu}_1$

$$\begin{aligned}
\mathbb{P}\{f_1(\overline{\mu})\} &= \mathbb{P}\left[\overline{\mu}_1(X) + \frac{A}{\pi}\Big\{Y - \overline{\mu}_1(X)\Big\}\right] \\
&= \mathbb{P}\left[\overline{\mu}_1(X) + \frac{\pi}{\pi}\Big\{\mu_1(X) - \overline{\mu}_1(X)\Big\}\right] \\
&= \mathbb{E}\{\mu_1(X)\} = \psi_1
\end{aligned} \tag{3.5}$$

where the second equality used iterated expectation and the Bernoulli randomization. Therefore we have

$$\mathbb{E}(\widehat{\psi}_1 \mid D^n) = \mathbb{P}\{f(\widehat{\mu}_1)\} = \psi_1$$

where the first equality uses the fact that $\widehat{\mu}_a(x)$ is fixed given independent D^n and the iid assumption, and the second (3.5). □

Theorem 3.2 is a simple but powerful result. It shows the doubly robust estimator is exactly unbiased, *for any choice of* regression estimator $\widehat{\mu}_a$. Hence, although the estimator $\widehat{\psi}$ exploits covariate information, its bias is not at all affected by accidentally misspecified models or biased regression estimators with slow convergence rates.

Remark 3.7. Theorem 3.2 also has an important implication for understanding the variance and limiting distribution of $\widehat{\psi}$. Namely, the logic in the proof shows that

$$\mathbb{P}\{f(\overline{\mu})\} = \psi$$

for *any* (fixed) $\overline{\mu}$. This means that, since $\widehat{\psi}$ is a sample average of an estimated function and thus the decomposition from Lemma 3.1 holds, we can write

$$\begin{aligned} \widehat{\psi} - \psi &= (\mathbb{P}_n - \mathbb{P})(\widehat{f} - \overline{f}) + \mathbb{P}(\widehat{f} - \overline{f}) + (\mathbb{P}_n - \mathbb{P})\overline{f} \\ &\equiv T_1 + T_2 + Z^* \end{aligned} \tag{3.6}$$

for *any* $\overline{f} = f(\overline{\mu})$. Since it will be useful in our analysis for $\widehat{f}$ to be consistent for $\overline{f}$, we will simply define $\overline{f} = f(\overline{\mu})$ to be the corresponding probability limit, i.e., by taking $\overline{\mu}_a$ to be a fixed function such that $\|\widehat{\mu}_a - \overline{\mu}_a\| = o_{\mathbb{P}}(1)$. We will see that this will allow us to completely sidestep whether the estimator $\widehat{\mu}_a$ is consistent for the *true* regression function μ_a, and instead just require that it be consistent for *something*.

Now we tackle the limiting distribution of $\widehat{\psi}$. Recall we know Z^* in the decomposition (3.6) is asymptotically normal, so we only need to understand the T_1 and T_2 terms. First we provide a general analysis of the first term

$$T_1 = (\mathbb{P}_n - \mathbb{P})(\widehat{f} - \overline{f})$$

in that decomposition.

Lemma 3.3. *Let $\mathbb{P}_n$ denote the empirical measure over $Z^n = (Z_1, ..., Z_n)$, and let $\widehat{f}(z)$ be any function estimated from a sample $D^N = (Z_{n+1}, ..., Z_{n+N})$, which is independent of Z^n. Then*

$$(\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - f\|}{\sqrt{n}}\right).$$

Proof. See Kennedy et al. [2019a]. □

Lemma 3.3 shows that T_1 terms are asymptotically negligible, i.e., that $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$, as long as $\widehat{f}$ is consistent for f (or $\overline{f}$ in our case, which will hold by definition).

The next result gives the limiting distribution of the doubly robust estimator, under no assumptions beyond the experiment design (and iid sampling) and that the regression estimators $\widehat{\mu}_a$ converge to anything at any rate.

Theorem 3.3. *Consider an iid Bernoulli experiment with $\mathbb{P}(A=1)=\pi$. Suppose the regression estimators $\widehat{\mu}_a$ are:*

1. *constructed from a separate independent sample, and*
2. *consistent (at any rate) for some functions $\overline{\mu}_a$ (not necessarily the true regression functions μ_a) in the sense that $\|\widehat{\mu}_a - \overline{\mu}_a\| = o_{\mathbb{P}}(1)$.*

Then the doubly robust estimator $\widehat{\psi}$ is root-n consistent and asymptotically normal with

$$\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N\Big(0, var(\overline{f})\Big)$$

where $\overline{f} = f(\overline{\mu})$ is defined as in (3.4).

Proof. By Lemma 3.1 we can write the decomposition (3.6) with $\overline{f} = f(\overline{\mu})$ for any $\overline{\mu}$. We will define $\overline{\mu}$ as the probability limit of $\widehat{\mu}$, as in the statement of the theorem.

By Lemma 3.3, we have $T_1 = O_{\mathbb{P}}(\|\widehat{f} - \overline{f}\|/\sqrt{n})$. Now note

$$\begin{aligned}
\|\widehat{f}_1 - \overline{f}_1\|^2 &= \left\| \widehat{\mu}_1 + \frac{A}{\pi}\Big\{Y - \widehat{\mu}_1(X)\Big\} - \overline{\mu}_1 - \frac{A}{\pi}\Big\{Y - \overline{\mu}_1(X)\Big\} \right\|^2 \\
&= \left\| \Big\{\widehat{\mu}_1 - \overline{\mu}_1\Big\} \left\{1 - \frac{A}{\pi}\right\} \right\|^2 \\
&= \int \Big\{\widehat{\mu}_1(x) - \overline{\mu}_1(x)\Big\}^2 \left(\frac{A-\pi}{\pi}\right)^2 \, d\mathbb{P}(z) \\
&= \left(\frac{\text{var}(A)}{\pi^2}\right) \int \Big\{\widehat{\mu}_1(x) - \overline{\mu}_1(x)\Big\}^2 \, d\mathbb{P}(x) \\
&= \left(\frac{1-\pi}{\pi}\right) \|\widehat{\mu}_1 - \overline{\mu}_1\|^2
\end{aligned}$$

where the fourth equality used the Bernoulli randomization. The same logic applies to $\|\widehat{f}_0 - \overline{f}_0\|$, and so by the triangle inequality

$$T_1 = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - \overline{f}\|}{\sqrt{n}}\right) = O_{\mathbb{P}}\left(\frac{\|\widehat{\mu}_1 - \overline{\mu}_1\| + \|\widehat{\mu}_0 - \overline{\mu}_0\|}{\sqrt{n}}\right)$$

which is $o_{\mathbb{P}}(1/\sqrt{n})$ since $\|\widehat{\mu}_a - \overline{\mu}_a\| = o_{\mathbb{P}}(1)$ by definition.

For the T_2 term, we have $\mathbb{P}(\widehat{f} - \overline{f}) = 0$ by (3.5). This gives the result. □

Theorem 3.3 shows that not only is the doubly robust estimator $\widehat{\psi}$ unbiased for any choice of regression estimator, it is also root-n consistent and asymptotically normal – even if the estimators $\widehat{\mu}_a$ are completely misspecified, and or converging at arbitrarily slow rates. This is a pretty amazing result.

This immediately implies that distribution-free confidence intervals can be constructed as in the following corollary.

Corollary 3.2. *Under the assumptions of Theorem 3.3, a distribution-free asymptotic 95% confidence interval for the average treatment effect ψ is given by*

$$\widehat{\psi} \pm 1.96\sqrt{\frac{\widehat{var}\{f(\widehat{\mu})\}}{n}}.$$

Further, finite-sample variance bounds can be constructed using the same logic as in the proof of Theorem 3.3.

Proposition 3.3. *Under the assumptions of Theorem 3.3, the doubly robust estimator $\widehat{\psi}$ in (3.3) has variance at most*

$$var(\widehat{\psi}) \leq \frac{1}{n}\left\{var(\overline{f}) + \left(\frac{1-\pi}{\pi}\right)\|\widehat{\mu}_1 - \overline{\mu}_1\|^2 + \left(\frac{\pi}{1-\pi}\right)\|\widehat{\mu}_0 - \overline{\mu}_0\|^2\right\}.$$

3.4.3 Efficiency

We have learned the surprising result that the sample-split doubly robust estimator is exactly unbiased for any choice of regression estimator $\widehat{\mu}_a$, and root-n consistent and asymptotically normal as long as $\widehat{\mu}_a$ converges to some fixed function at any rate. As would be expected, the efficiency of the doubly robust estimator depends on the probability limits $\overline{\mu}_a$ that the regression estimators $\widehat{\mu}_a$ converge to. This raises some important questions:

- What is the best possible (i.e., most efficient) probability limit $\overline{\mu}_a$?
- Is the doubly robust estimator necessarily more efficient than the difference-in-means or Horvitz-Thompson estimator?

Recall however that the difference-in-means and Horvitz-Thompson estimators can be written as variants of the doubly robust estimator, for particular choices of $\widehat{\mu}_a$. Therefore the best choice of $\overline{\mu}_a$ will dominate others in this class.

The next result shows what you might expect: that the best limit $\overline{\mu}_a$ in terms of efficiency is the *true* regression function μ_a (recall this limit is irrelevant for bias since $\widehat{\psi}$ is unbiased for any $\widehat{\mu}_a$).

Theorem 3.4. *Define $f(\overline{\mu})$ as in (3.4). Then for any $\overline{\mu} = (\overline{\mu}_1, \overline{\mu}_0)$ with $\overline{\mu}_a : \mathcal{X} \mapsto \mathbb{R}$*

$$var\{f(\overline{\mu})\} \geq var\{f(\mu)\}$$

where $\mu = (\mu_1, \mu_0)$ denotes the true regression functions.

Proof. We have

$$\begin{aligned}
\mathrm{var}\{f(\overline{\mu})\} &= \mathrm{var}\left[\overline{\mu}_1(X) - \overline{\mu}_0(X) + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \overline{\mu}_A(X)\right\}\right] \\
&= \mathrm{var}\Bigg\{(\mu_1 - \mu_0) + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)(Y - \mu_A) \\
&\qquad + \left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\Bigg\} \\
&= \mathrm{var}\{f(\mu)\} + \mathrm{var}\left\{\left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\} \\
&\qquad + 2\mathrm{cov}\left\{f(\mu), \left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\}
\end{aligned}$$

But the latter covariance is zero since

$$\begin{aligned}
&\mathrm{cov}\left\{f(\mu), \left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\} \\
&\qquad = \mathbb{E}\left[(\mu_1 - \mu_0 - \psi)\left\{\left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\}\right] \\
&\qquad = 0
\end{aligned}$$

where the second equality follows from iterated expectation since $\mathbb{E}\{f(\mu) \mid X, A\} = \mu_1 - \mu_0$, and the third since $A \perp\!\!\!\perp X$ so that $\mathbb{E}\{Ag(X)\} = \pi\mathbb{E}\{g(X)\}$ for any g. This gives the result □

Theorem 3.4 is critically informative about how to construct the doubly robust estimator $\widehat{\psi}$ in practice. Namely, it indicates that we should estimate the regression functions as flexibly as possible: bias is zero regardless, and efficiency is optimized when the regression functions are estimated consistently. This is a special case not often seen in statistics where there is essentially no penalty (at least asymptotically) for slow rates of convergence, and important benefits for consistency.

However the second question still remains: when based on a misspecified model for μ_a, does the doubly robust estimator necessarily still improve efficiency (say relative to the Horvitz-Thompson estimator)? In fact, this is not necessarily so, for particularly misspecified choices of $\widehat{\mu}_a$. However, there are multiple approaches that can be used to guarantee efficiency gains. One simple option proposed by Rubin and van der Laan [2008] is to posit a working parametric model $\mu_a(x) = \mu_a(x;\beta)$, but rather than estimating the parameters via maximum likelihood, instead estimate parameters by picking those that minimize an estimator of the variance, i.e., use

$$\widetilde{\beta} = \arg\min_{\beta} \ \widehat{\mathrm{var}}\left[\mu_1(X;\beta) - \mu_0(X;\beta) + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \mu_A(X;\beta)\right\}\right].$$

Other approaches similar in spirit are also possible [Tan, 2010].

3.4.4 Back to the Plug-In

In the previous section we saw strong evidence that, if one wants to remain agnostic about the data-generating process beyond the known randomization probabilities, retaining robustness while exploiting covariate information to gain efficiency, then the doubly robust estimator (3.3) using a flexible regression estimator is a good choice. In particular, it will be root-n consistent and asymptotically normal as long as the regression estimator $\widehat{\mu}_a$ converges to *anything at any rate*, and if the regression estimator $\widehat{\mu}_a$ is consistent for the true regression function (again *at any rate*) then it will be asymptotically efficient.

However, in practice, applied researchers often use ordinary least squares or plug-in estimators based on parametric models. Do we have any basis for trusting such results? In fact, the following surprising result shows that some if not many parametric plug-in estimators can be represented in the doubly robust form: they are doubly robust estimators disguised as plug-ins. (Though it is important to note that this is not true of all plug-in estimators.)

Proposition 3.4. *Suppose regression predictions $\widehat{\mu}_a$ satisfy*

$$\mathbb{P}_n\left[(1, A)^{\mathrm{T}}\left\{Y - \widehat{\mu}_A(X)\right\}\right] = 0 \tag{3.7}$$

where $A \perp\!\!\!\perp X$ is randomized according to a Bernoulli experiment. Then the parametric plug-in estimator

$$\mathbb{P}_n\left\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\right\}$$

is numerically equivalent to the doubly robust estimator

$$\mathbb{P}_n\left[\left\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\right\} + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \widehat{\mu}_A(X)\right\}\right].$$

Proof. Since

$$\mathbb{P}_n\left[(1, A)^{\mathrm{T}}\left\{Y - \widehat{\mu}_A(X)\right\}\right] = 0$$

it follows that

$$\frac{1}{\pi}\mathbb{P}_n\left[A\left\{Y - \widehat{\mu}_A(X)\right\}\right] = \frac{1}{1-\pi}\mathbb{P}_n\left[A\left\{Y - \widehat{\mu}_A(X)\right\}\right] = \frac{1}{1-\pi}\mathbb{P}_n\left[\left\{Y - \widehat{\mu}_A(X)\right\}\right] = 0.$$

Therefore

$$\mathbb{P}_n\left[\left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \widehat{\mu}_A(X)\right\}\right] = 0$$

so that the correction term in the doubly robust zero is estimator, and the plug-in and doubly robust estimator are equal. □

The sufficient condition (3.7) in Proposition 3.4 says that the $\widehat{\mu}_a$ residuals must average to zero both in the whole sample and among the treated. This will hold for example in generalized linear models with an intercept and main effect term for treatment. Thus Proposition 3.4 shows that, although our earlier analysis of the parametric plug-in appeared to hinge on restrictive parametric model assumptions, this is not necessarily so – at least in Bernoulli experiments, and for plug-ins based on models with an intercept and main effect for treatment. Such parametric plug-in estimators will be root-n consistent for the average treatment effect (and asymptotically normal), even under misspecification of the regression estimator $\widehat{\mu}_a$, as long as it converges in probability to anything at any rate (a very weak condition).

Remark 3.8. Although a plug-in whose regression estimates satisfy (3.7) will take on all the advantageous robustness and efficiency properties of the doubly robust estimator, note that variance estimates must be based on the doubly robust variance as in Corollary 3.2. Otherwise (e.g., if based on the parametric model being correct as in Theorem 3.1) corresponding confidence intervals and hypothesis tests may not be valid.

Remark 3.9. The condition (3.7) holding will generally not be enough to ensure that a plug-in will be doubly robust, outside of a Bernoulli experiment where treatment is completely independent of covariates. For example, this would not be sufficient in the conditionally randomized experiment discussed next, since the randomization probabilities cannot be brought outside the average as in the proof of Proposition 3.4.

3.5 Conditional Randomization

In conditionally randomized experiments, the randomization probabilities can differ by covariate values, e.g., in a stratified Bernoulli experiment one sets

$$\mathbb{P}(A = 1 \mid X = x, Y^a) = \pi(x)$$

where the function $\pi(x)$ can vary with x (recall $\pi(x) = \pi$ in a Bernoulli experiment).

Experiments may use conditional or stratified randomization to improve efficiency (e.g., by treating more units at covariate values where treated outcomes are more variable than control outcomes), or to improve subject outcomes (e.g., by treating more units at covariate values where treated outcomes are likely to be higher than control outcomes, and treating fewer when treatment is ineffective or even harmful).

Doubly robust estimators take the same form as (3.3), but replace π with $\pi(x)$, i.e.,

$$\mathbb{P}_n \left[\left\{ \widehat{\mu}_1(X) - \widehat{\mu}_0(X) \right\} + \left\{ \frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)} \right\} \left\{ Y - \widehat{\mu}_A(X) \right\} \right]$$

and the logic of the theoretical analysis is the same as well.

There are some important differences between simple Bernoulli experiments and conditionally randomized designs, however. First, the difference-in-means estimator is no longer a valid estimator, since it is no longer the case that $A \perp\!\!\!\perp Y^a$ or $A \perp\!\!\!\perp (X, Y^a)$; instead, in a conditionally randomized experiment it only holds that $A \perp\!\!\!\perp Y^a \mid X$. Second, plug-in estimators are not in general doubly robust in conditionally randomized designs, even when they satisfy the condition (3.7); this is because the randomization probabilities cannot be brought outside the average as in the proof of Proposition 3.4.

Chapter 4

Unconfounded Observational Studies

We have learned that experiments allow for efficient and unbiased causal inference, through controlled (randomized) treatment assignment. However, for many problems experiments would be impossible to implement, or unethical, or too costly. For example we cannot randomize people to smoke, or become obese; an experiment studying lifetime diets on mortality would take decades (and be riddled with noncompliance). Further, detailed observational data is often collected and readily available: what should be done with such data?

In some ways observational studies can resemble experiments. For example the observed data structure is the same – we still observe a triplet of covariates, treatment, and outcome $(X, A, Y) \sim \mathbb{P}$. We will also continue to assume the consistency condition ($Y = Y^a$ if $A = a$), so that the observed outcome equals the potential outcome under the observed treatment.

However there is a major underlying difference: in observational studies, the treatment happened "naturally" according to some unknown process, and was not under experimenters' control. Consider some examples. Cancer patients may decide to have surgery or not based on myriad factors: current health, past medical history, conversations with doctors and family, assessment of risk-benefit trade-offs, etc. Class size is not random and instead depends on popularity of subject matter, quality of lecturer, etc. Gun laws vary widely across US states; they are most restrictive in northeast states, and least restrictive in northwest and southeast states. Reasons for this variation might include, for example, cultural differences or reactions to specific shootings or events (e.g., bans on assault weapons and bump stocks were introduced after shootings in 2012 at Sandy Hook Elementary School and in 2017 in Las Vegas, respectively).

Mathematically, this means that in observational studies the treatment distribution

$$\mathbb{P}(A = a \mid X, Y^{a'})$$

is unknown. In contrast, recall that in Bernoulli experiments it is known by design that $\mathbb{P}(A = 1 \mid X, Y^a) = \mathbb{P}(A = 1) = \pi$.

4.1 No Identification Without Assumptions

When the treatment A and potential outcomes Y^a can be correlated, then there is unfortunately no hope for identification. Intuitively this is because consistency only lets us learn about potential outcomes under treatment among those who were actually treated – those who were not may be arbitrarily different.

Formally we can write the mean potential outcome under treatment as

$$\begin{aligned}\mathbb{E}(Y^1) &= \mathbb{E}(Y^1 \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y^1 \mid A=1)\mathbb{P}(A=1) \\ &= \mathbb{E}(Y^1 \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y \mid A=1)\mathbb{P}(A=1)\end{aligned}$$

where the first equality follows by the law of total probability, and the second by consistency. However in general the observed data distribution $\mathbb{P}$ says nothing about the quantity $\mathbb{E}(Y^1 \mid A=0)$. In general, $\mathbb{E}(Y^1 \mid X, A=0) \neq \mathbb{E}(Y^1 \mid X, A=1)$ since those who take control may be completely different from those who take treatment.

The next proposition shows that, if only relying on consistency, one can merely bound rather than point identify mean potential outcomes; further, these bounds are necessarily imprecise and cannot identify whether the treatment has a non-zero effect.

Proposition 4.1. *Let $(A, Y) \sim \mathbb{P}$ with $\mathbb{P}(Y \in [0,1]) = 1$ and assume consistency so that $Y = Y^a$ if $A = a$. Then*

$$\mathbb{E}\{(2A-1)Y\} - \mathbb{P}(A=1) \leq \mathbb{E}(Y^1 - Y^0) \leq \mathbb{E}\{(2A-1)Y\} + \mathbb{P}(A=0)$$

and these bounds are sharp.

Proof. We have

$$\begin{aligned}\mathbb{E}(Y^1) &= \mathbb{E}(Y^1 \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y^1 \mid A=1)\mathbb{P}(A=1) \\ &= \mathbb{E}(Y^1 \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y \mid A=1)\mathbb{P}(A=1) \\ &\in \Big[\mathbb{E}(Y \mid A=1)\mathbb{P}(A=1),\ \mathbb{P}(A=0) + \mathbb{E}(Y \mid A=1)\mathbb{P}(A=1)\Big] \\ &= \Big[\mathbb{E}(AY),\ \mathbb{P}(A=0) + \mathbb{E}(AY)\Big]\end{aligned}$$

where the first equality follows by the law of total probability, the second by consistency, and the last bounds by the fact that $Y \in [0,1]$. By the same logic

$$\begin{aligned}\mathbb{E}(Y^0) &= \mathbb{E}(Y^0 \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y^0 \mid A=1)\mathbb{P}(A=1) \\ &= \mathbb{E}(Y \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y^0 \mid A=1)\mathbb{P}(A=1) \\ &\in \Big[\mathbb{E}\{Y(1-A)\},\ \mathbb{E}\{Y(1-A)\} + \mathbb{P}(A=1)\Big]\end{aligned}$$

Taking the difference of the lower and upper bounds (and vice versa) yields the result. The lower bound is attained when $Y^a = 1-a$ among those with $A = 1-a$, and the upper bound when $Y^a = a$ among those with $A = 1-a$. □

Remark 4.1. The assumption that $Y \in [0,1]$ with probability one is immaterial as long as Y has bounded support, since any $Y \in [a,b]$ can always be rescaled as $\frac{Y-a}{b-a} \in [0,1]$.

Note that the length of the above bounds is exactly one, so they must necessarily include zero. This implies that without further assumptions it is impossible to rule out whether a treatment has no effect. This should make you question any claim of assumption-free causal inference, or general test for unmeasured confounding.

4.2 Identification via Confounder Measurement

Without any assumptions beyond consistency, we have seen that is impossible to rule out the possibility of zero treatment effect, even with infinite data, and that the value of the treatment effect can at best be bounded. One way to make progress is to try to collect as many relevant covariates X as possible to be able to explain the treatment process, in the sense that

$$A \perp\!\!\!\perp Y^a \mid X. \tag{4.1}$$

Condition (4.1) goes by several names in the literature: exchangeability, ignorability, or no unmeasured confounding. It means treatment is essentially randomized within levels of the covariates, since it is conditionally independent of potential outcomes. In other words, there can be no remaining unmeasured confounders U that may induce a correlation between the treatment and potential outcomes. Condition (4.1) can be illustrated graphically as in the directed acyclic graph in Figure 4.1.

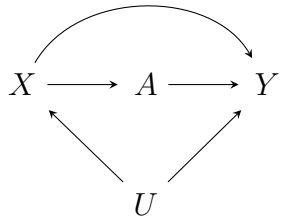


Figure 4.1: A directed acyclic graph representation of the no unmeasured confounding condition (4.1), which holds due to the missing arrow $U \to A$.

Importantly, no unmeasured confounding as in (4.1) means the observational study is actually a conditionally randomized experiment, but one in which the randomization probabilities

$$\mathbb{P}(A = a \mid X = x, Y^{a'} = u) = \mathbb{P}(A = a \mid X = x)$$

are unknown. In other words, an unconfounded observational study can be viewed as a setup where nature ran a conditionally randomized experiment, but kept the randomization probabilities hidden from us.

A critical question is: how does one verify or justify the assumption (4.1)?

Condition (4.1) is generally impossible to test with data, since Y^a is not directly observed (under consistency Y^a is only observed among those with $A = a$, so one cannot measure the correlation between A and Y^a). This means that causal inference cannot be purely data driven – some subject matter knowledge is required. In fact, (4.1) can generally only be justified non-mathematically using subject matter expertise. For example, one might try to understand how physicians are assigning treatment, or how people select into job training programs, etc.

Remark 4.2. The no unmeasured confounding condition (4.1) is commonly invoked, but it is by no means the only assumption or strategy one might consider for causal inference in non-experimental settings. In future chapters we will study alternatives.

Recall that (4.1) also held in the Bernoulli experiments we considered; however there are some important differences. First, in an unconfounded observational study we only have $A \perp\!\!\!\perp Y^a \mid X$, and not $A \perp\!\!\!\perp (X, Y^a)$. This means that there are important differences among subjects receiving different treatment levels (but that these differences are only relevant insofar as they appear in observed covariates). Second, the treatment distribution $\mathbb{P}(A = a \mid X = x)$ is unknown and thus would have to be estimated. Third, we can never really be sure that (4.1) holds in an observational study, whereas in an experiment we know with certainty that it holds due to the design.

Remark 4.3. When treatment is binary, the quantity $\mathbb{P}(A = 1 \mid X = x) = \pi(x)$ is known as the "propensity score". It represents the conditional chance of receiving treatment for subjects with covariates $X = x$.

The strategy for identifying average treatment effects in unconfounded observational studies is essentially the same as that used in experiments. This is formalized in the next proposition.

Proposition 4.2. *Let* $(X, A, Y) \sim \mathbb{P}$ *and assume:*

1. *Consistency:* $Y = Y^a$ *if* $A = a$.
2. *No unmeasured confounding:* $A \perp\!\!\!\perp Y^a \mid X$.
3. *Positivity:* $\mathbb{P}(A = a \mid X = x) > 0$ *with probability one.*

Then

$$\mathbb{E}(Y^a) = \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}.$$

Proof. We have

$$\begin{aligned}\mathbb{E}(Y^a) &= \mathbb{E}\{\mathbb{E}(Y^a \mid X)\} = \mathbb{E}\{\mathbb{E}(Y^a \mid X, A = a)\} \\ &= \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}\end{aligned}$$

where the first equality follows by iterated expectation, the second by no unmeasured confounding, and the third by consistency. Positivity is required so that the conditional expectations are well-defined. □

In Proposition 4.2 we needed a *positivity* condition, which was implicit in experiments since it held by design. Namely positivity requires that the propensity score be bounded away from extreme values; in other words, to estimate the mean potential outcome if everyone were treated at level $A = a$, we require that everyone has some non-zero chance of being treated at that level. In experiments this holds as long as the proverbial coins that are flipped to decide treatment assignment are not deterministic. However, in observational studies, positivity is a real concern; some subjects may really have zero chance of receiving treatments other than the one they actually received. For example, very sick patients may never not receive treatment, or very healthy patients may never have intensive life-saving surgeries.

Positivity is sometimes called the "experimental treatment assumption" or "overlap". And it can be stated in various more or less equivalent ways. For example, with binary treatments the following are equivalent:

- $0 < \pi(x) < 1$ for all $x \in \mathcal{X}$ with $\mathbb{P}(X = x) > 0$
- $\mathbb{P}\{0 < \pi(X) < 1\} = 1$
- $0 < d\mathbb{P}(x \mid A = 1)/d\mathbb{P}(x \mid A = 0) < \infty$

Positivity is also sometimes written as

- $\mathbb{P}\{\epsilon \leq \pi(X) \leq 1 - \epsilon\} = 1$ for some $\epsilon > 0$

The first two forms are sufficient for identification, while the third is often used for analyzing estimators, since arbitrarily poor performance may be possible if the propensity scores can be arbitrarily close to zero (even if positive). For simplicity, we will often just use the stronger ϵ bound.

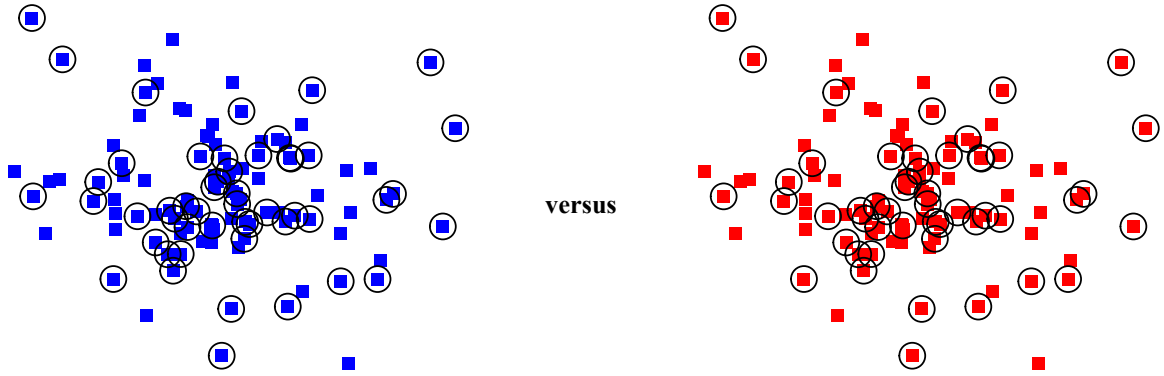
4.2.1 Effects of Treatment on the Treated

So far we have focused on identification of average treatment effects, e.g., of the form $\mathbb{E}(Y^1 - Y^0)$. However, it is also common (especially in observational studies) to pursue the average treatment effect on the treated given by

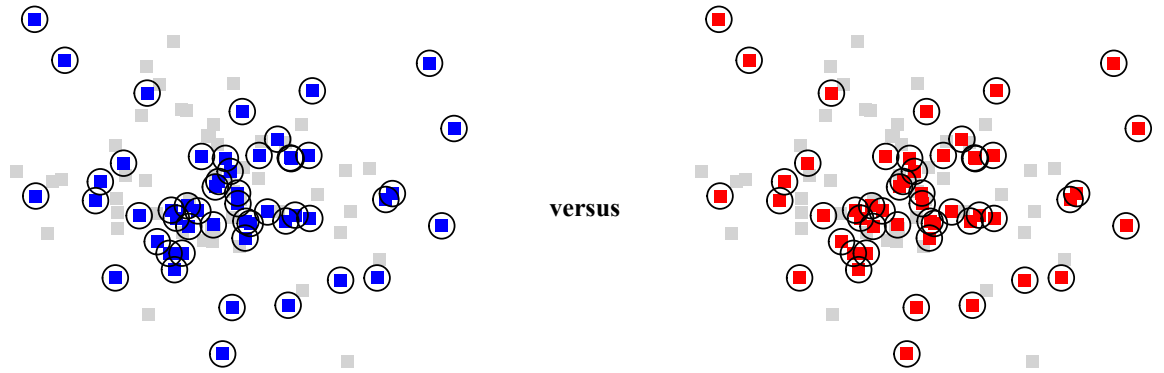
$$\psi_{att} = \mathbb{E}(Y^1 - Y^0 \mid A = 1) = \mathbb{E}(Y - Y^0 \mid A = 1)$$

In contrast to the average treatment effect $\mathbb{E}(Y^1 - Y^0)$ the effect of treatment on the treated measures the difference in mean outcomes if treatment was removed from those who received it. This parameter, and its relation to the average effect, is illustrated in the schematic given in Figure 4.2.

The effect on the treated parameter can be useful if the goal is to learn effects of removing an exposure, e.g., when assigning everyone in the population treatment may not be feasible. It also requires weaker identifying assumptions, as illustrated in the next proposition.



(a) Average treatment effect



(b) Effect on the treated

Figure 4.2: Illustration of average treatment effect versus effect on the treated parameters. Each point in the plot represents a subject in some population. Circled points represent subjects who actually received treatment, and the color of the point represents the treatment they would receive based on the counterfactual contrast of interest, with blue denoting treatment and red control (and gray meaning they are excluded). The average treatment effect in (a) is a comparison of the mean outcome if all subjects were treated (regardless of actual treatment) versus if none were treated. In contrast, the effect on the treated in (b) compares the mean outcome among those actually treated to what it would have been if treatment was removed (among only this subpopulation).

Proposition 4.3. *Let* $(X, A, Y) \sim \mathbb{P}$ *and assume:*

1. *Consistency:* $Y = Y^a$ *if* $A = a$.
2. *No unmeasured confounding:* $A \perp\!\!\!\perp Y^0 \mid X$.
3. *Positivity:* $\mathbb{P}(A = 0 \mid X = x) > 0$ *with probability one among those with* $A = 1$.

Then

$$\mathbb{E}(Y^1 - Y^0 \mid A = 1) = \mathbb{E}\{Y - \mathbb{E}(Y \mid X, A = 0) \mid A = 1\}.$$

Proof. By consistency, it follows that $\mathbb{E}(Y^1 \mid A = 1) = \mathbb{E}(Y \mid A = 1)$. Then

$$\begin{aligned}\mathbb{E}(Y^0 \mid A = 1) &= \mathbb{E}\{\mathbb{E}(Y^0 \mid X, A = 1) \mid A = 1\} \\ &= \mathbb{E}\{\mathbb{E}(Y^0 \mid X, A = 0) \mid A = 1\} \\ &= \mathbb{E}\{\mathbb{E}(Y \mid X, A = 0) \mid A = 1\}\end{aligned}$$

where the first equality follows by iterated expectation, the second by no unmeasured confounding, and the third by consistency. Positivity is required so that the conditional expectations and their averages are well-defined. □

Intuitively, since the first quantity $\mathbb{E}(Y^1 \mid A = 1) = \mathbb{E}(Y \mid A = 1)$ in the effect on the treated is identified under no conditions, one only needs $A \perp\!\!\!\perp Y^0 \mid X$ and similarly "one-sided" positivity, in that the propensity scores only need to be bounded away from one. The latter follows since anyone with $\pi(x) = 0$ will necessarily not be part of the $A = 1$ population, and thus there is no need to assess their outcome had they been treated. On the other hand, if there are subjects with $\pi(x) = 1$ then they would be part of the $A = 1$ population, but learning about their outcome under control would be impossible.

4.3 Observational Studies versus Experiments

The identification results from Propositions 4.2 and 4.3 have converted causal problems into purely statistical ones. For example, in the average treatment effect case, after identification the goal has become to estimate the averaged regression function

$$\mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}$$

as well as possible; this is a purely statistical problem. The difficulty compared to the experimental setup is that in an observational study the treatment distribution is unknown. Before we move to estimation, however, we will discuss the experimental/observational distinction in some more detail.

A fascinating debate can be had about the evidential status of observational studies versus experiments. There are extremists on both sides, but the reality is more nuanced and context-dependent. The main issues stem from validity, feasibility, and generalizability.

Here are some sentiments you might hear from a "Pro-Observational Extremist":

- Causality is easy: just throw basic covariates into a regression model and voila!
- We should not waste time and money on experiments, when we can just do cheap and easy observational studies.

These sentiments are sometimes implicit rather than explicit, but examples abound, for example in exploratory research with catchy titles meant to make headlines.

Here are some sentiments you might hear from a "Pro-Experimental Extremist":

- The only way to learn anything reliable is with the gold standard: experiments.
- All experiments are necessarily trustworthy and informative, and anything else is not real science.

Examples of this kind of extremism can be found among defenses of tobacco and pharmaceutical companies. For example Michaels [2008] recounts a cigarette executive stating that "Doubt is our product, since it is the best means of competing with the 'body of fact' that exists in the minds of the general public." Statisticians are probably more likely to be pro-experimental extremists than pro-observational extremists.

As is often the case with extremist perspectives, real life is more nuanced.

Here are some issues with the pro-observational extremist perspective:

- Causal claims from observational studies require untestable assumptions that can be difficult to assess.
- Adjusting for measured covariates may not be enough, since there could always be some unmeasured confounding that was missed.
- Even if all relevant confounders happened to be measured, appropriate adjustment and valid inference is non-trivial in nonparametric or high-dimensional models.

Here are some issues with the pro-experimental extremist perspective:

- Experiments are not always feasible (e.g.,studies of long-term effects) or ethical (e.g., smoking).
- Experiments are often conducted in selected, non-representative populations: an unbiased estimate in the wrong population may be less useful than a slightly biased estimate in the right one.
- Experiments are often plagued with non-compliance and missing data, which can essentially turn them into observational studies.

In sum, observational studies and experiments are not necessarily uniformly bad or good; both types of studies range in quality, and it is not true that one type dominates the other. Their evidential status is context-dependent and needs to be evaluated case-by-case, based on specific merits or faults.

4.4 Estimation of Average Effects

In causal inference problems, one can often categorize methods for estimating treatment effects as being based on regression, weighting, or both (doubly robust).

For the average treatment effect $\psi = \mathbb{E}(Y^1 - Y^0)$, regression estimators can be motivated based on the identifying expression

$$\psi = \mathbb{E}\Big\{\mathbb{E}(Y^1 \mid X, A = 1) - \mathbb{E}(Y^0 \mid X, A = 0)\Big\} = \mathbb{E}\Big\{\mu_1(X) - \mu_0(X)\Big\}$$

which suggests the regression estimator

$$\widehat{\psi}_{reg} = \mathbb{P}_n\Big\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\Big\}. \tag{4.2}$$

Operationally, this estimator predicts (the conditional mean of) the potential outcomes Y^1 and Y^0 for each subject, takes the difference, and averages across the sample. One can also interpret this estimator with reference to matching: for each unit with a particular $X = x$ value, one finds unit(s) with the same or similar $X = x$ value but who received the opposite treatment.

Weighting estimators can be motivated based on the inverse-probability-weighted expression

$$\psi = \mathbb{E}\left[\left\{\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right\}Y\right] = \mathbb{E}\Big\{\mu_1(X) - \mu_0(X)\Big\}$$

which suggests the inverse-probability-weighted estimator

$$\widehat{\psi}_{ipw} = \mathbb{P}_n\left[\left\{\frac{A}{\widehat{\pi}(X)} - \frac{1-A}{1-\widehat{\pi}(X)}\right\}Y\right]. \tag{4.3}$$

This estimator can be viewed as up- or down-weighting observations whose covariates are under- or over-represented in their treated group compared to the population covariate distribution. This is similar in spirit to importance sampling: the covariate distribution for the treated is different from that in the general population, so one needs to use a change of measure to reweight treated outcomes appropriately. It is also popular to view the inverse weighting as creating a "pseudopopulation" of treated units whose covariate distribution matches that of the entire population.

Doubly robust estimators can be motivated based on the expressions

$$\begin{aligned}\psi &= \mathbb{E}\left[\left\{\frac{A}{\overline{\pi}(X)} - \frac{1-A}{1-\overline{\pi}(X)}\right\}\Big\{Y - \mu_A(X)\Big\} + \Big\{\mu_1(X) - \mu_0(X)\Big\}\right] \\ &= \mathbb{E}\left[\left\{\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right\}\Big\{Y - \overline{\mu}_A(X)\Big\} + \Big\{\overline{\mu}_1(X) - \overline{\mu}_0(X)\Big\}\right]\end{aligned}$$

which hold for any $(\overline{\pi}, \overline{\mu})$. This suggests the estimator

$$\widehat{\psi}_{dr} = \mathbb{P}_n \left[\left\{ \frac{A}{\widehat{\pi}(X)} - \frac{1-A}{1-\widehat{\pi}(X)} \right\} \left\{ Y - \widehat{\mu}_A(X) \right\} + \left\{ \widehat{\mu}_1(X) - \widehat{\mu}_0(X) \right\} \right] \tag{4.4}$$

which was also used in the previous chapter with experiments, except now the propensity score $\pi(x)$ depends on covariates and is unknown so needs to be estimated. The doubly robust estimator is somewhat less intuitive than the other two options, but it can be viewed as correcting leftover smoothing bias of a regression of inverse-probability-weighted estimator, or augmenting an inverse-probability-weighted estimator with regression predictions to increase efficiency. We will see in later chapters that its precise form comes from a bias correction based on a distributional Taylor expansion of the average treatment effect functional.

4.4.1 Discrete Covariates

For some intuition we will first consider the simplest case, where the covariates X are discrete and low-dimensional, i.e., $X \in \{1, ..., d\}$ with d fixed. We will see that in this setup, when one uses the empirical distribution to estimate the "nuisance functions" π and μ_a, then all three of the previously mentioned estimators coincide in that they are numerically equivalent. (Later we will show that they are asymptotically efficient in a local minimax sense). This numerical equivalence does not occur when the covariates have some continuous components and modeling or smoothing is used to construct the $\widehat{\pi}$ and $\widehat{\mu}_a$ estimates. Intuitively, the reason why all three estimators are numerically equivalent is because, when the covariates are discrete, there is no smoothness or additional structure to exploit, so each estimator is making full equivalent use of the data. Another way to think about it is that, in the discrete case, the empirical measure $\mathbb{P}_n$ is an actual valid distribution (including all conditional distributions), and so the identifying expression equalities above also hold for $\mathbb{P}_n$.

Our first result shows the numerical equivalence between the regression, weighting, and doubly robust estimators.

Proposition 4.4. *Suppose $X \in \{1, ..., d\}$ is discrete and the nuisance estimators are the empirical averages*

$$\widehat{\pi}(x) = \mathbb{P}_n(A \mid X = x) = \frac{\mathbb{P}_n\{A\mathbb{1}(X = x)\}}{\mathbb{P}_n\{\mathbb{1}(X = x)\}}$$
$$\widehat{\mu}_a(x) = \mathbb{P}_n(Y \mid X = x, A = a) = \frac{\mathbb{P}_n\{Y\mathbb{1}(A = a)\mathbb{1}(X = x)\}}{\mathbb{P}_n\{\mathbb{1}(A = a)\mathbb{1}(X = x)\}}$$

Then the regression, weighting, and doubly robust estimators defined in (4.2)–(4.4) are all numerically equivalent, i.e.,

$$\widehat{\psi}_{reg} = \widehat{\psi}_{ipw} = \widehat{\psi}_{dr}.$$

Proof. We will consider the $\psi_1 = \mathbb{E}\{\mu_1(X)\}$ term, since the logic is the same for ψ_0. To see that $\widehat{\psi}_{reg} = \widehat{\psi}_{ipw}$ note that

$$\begin{aligned}
\widehat{\psi}_{reg} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_1(X_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}_n\{YA\mathbb{1}(X=x_i)\}}{\mathbb{P}_n\{A\mathbb{1}(X=x_i)\}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\frac{1}{n}\sum_j Y_jA_j\mathbb{1}(X_j=x_i)}{\frac{1}{n}\sum_k A_k\mathbb{1}(X_k=x_i)\}} = \frac{1}{n}\sum_{j=1}^{n}Y_jA_j\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}(X_j=x_i)}{\frac{1}{n}\sum_k A_k\mathbb{1}(X_k=x_i)} \\
&= \frac{1}{n}\sum_{j=1}^{n}Y_jA_j\frac{\frac{1}{n}\sum_i \mathbb{1}(X_j=x_i)}{\frac{1}{n}\sum_k A_k\mathbb{1}(X_k=x_j)} = \frac{1}{n}\sum_{j=1}^{n}Y_jA_j/\widehat{\pi}(X_j) = \mathbb{P}_n\left\{\frac{AY}{\widehat{\pi}(X)}\right\} = \widehat{\psi}_{ipw}
\end{aligned}$$

where in the fifth equality we replace the x_i in the denominator with x_j since the numerator includes the indicator $\mathbb{1}(X_j = x_i)$.

Now to see that $\widehat{\psi}_{reg} = \widehat{\psi}_{dr}$ we will show $\mathbb{P}_n\{AY/\widehat{\pi}(X)\} = \mathbb{P}_n\{A\widehat{\mu}_1(X)/\widehat{\pi}(X)\}$, so that the correction term $\mathbb{P}_n[A\{Y-\widehat{\mu}_1(X)\}/\widehat{\pi}(X)] = 0$. Note

$$\begin{aligned}
\mathbb{P}_n\left\{\frac{A\widehat{\mu}_1(X)}{\widehat{\pi}(X)}\right\} &= \frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\widehat{\pi}(X_i)}\frac{\frac{1}{n}\sum_j Y_jA_j\mathbb{1}(X_j=x_i)}{\frac{1}{n}\sum_k A_k\mathbb{1}(X_k=x_i)} \\
&= \frac{1}{n}\sum_{j=1}^{n}Y_jA_j\frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\widehat{\pi}(X_i)}\frac{\mathbb{1}(X_j=x_i)}{\frac{1}{n}\sum_k A_k\mathbb{1}(X_k=x_i)} \\
&= \frac{1}{n}\sum_{j=1}^{n}Y_jA_j\frac{1}{\widehat{\pi}(X_j)}\frac{1}{n}\sum_{i=1}^{n}A_i\frac{\mathbb{1}(X_j=x_i)}{\frac{1}{n}\sum_k A_k\mathbb{1}(X_k=x_j)} \\
&= \frac{1}{n}\sum_{j=1}^{n}\frac{Y_jA_j}{\widehat{\pi}(X_j)} = \mathbb{P}_n\left\{\frac{AY}{\widehat{\pi}(X)}\right\}
\end{aligned}$$

where again in the third equality we replace x_i in the denominator with x_j due to the numerator indicator. This gives the result. □

Next we derive the limiting distribution of the estimator $\widehat{\psi}_{reg} = \widehat{\psi}_{ipw} = \widehat{\psi}_{dr}$.

Theorem 4.1. *Suppose $X \in \{1, ..., d\}$ is discrete and the nuisance estimators are the empirical averages from Proposition 4.4. Assume that Y is bounded and that $\pi(x)$ and $\widehat{\pi}(x)$ are bounded away from ϵ and $1-\epsilon$ for some $\epsilon > 0$ and all x. Then*

$$\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N(0, var(f))$$

for $\widehat{\psi}$ the estimators in (4.2)–(4.4) and

$$f(Z) = \mu_1(X) - \mu_0(X) + \left\{\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right\}\Big\{Y - \mu_A(X)\Big\}.$$

Proof. We will work with the $\widehat{\psi}_{dr}$ version of the estimator, which can be written as $\widehat{\psi}_{dr} = \mathbb{P}_n(\widehat{f})$ for

$$f(Z) = \mu_1(X) - \mu_0(X) + \left\{ \frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)} \right\} \Big\{ Y - \mu_A(X) \Big\}$$

and $\widehat{f}$ the version of f replacing (π, μ_a) with $(\widehat{\pi}, \widehat{\mu}_a)$.

Therefore by Lemma 3.1 we have the decomposition

$$\widehat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P})f + (\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) + \mathbb{P}(\widehat{f} - f) \equiv Z^* + T_1 + T_2.$$

We will first handle the T_1 term. Note since X is discrete we can write the nuisance estimators $(\widehat{\pi}, \widehat{\mu}_a)$ as linear regression estimators based on saturated models, i.e.,

$$\widehat{\pi}(x) = \pi(x; \widehat{\alpha}) = \widehat{\alpha}^{\mathrm{T}} w$$

where $w^{\mathrm{T}} = \{\mathbb{1}(x = 1), ..., \mathbb{1}(x = d-1)\} \in \{0,1\}^{d-1}$ and similarly

$$\widehat{\mu}_a(x) = \mu_a(x; \widehat{\beta}_a) = \widehat{\beta}_a^{\mathrm{T}} w.$$

This implies

$$|\widehat{f}(z) - f(z)| = |f(z; \widehat{\eta}) - f(z; \eta)| \leq C \|\widehat{\eta} - \eta\|$$

for $\eta = (\alpha, \beta_0, \beta_1)$ and $C < \infty$ some constant. Therefore f and $\widehat{f}$ belong to a Donsker class, which together with the central limit theorem and Lemma 19.24 from van der Vaart [2000] imply that $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$.

For the T_2 term, note that $f = f_1 - f_0$ for $f_a = \mu_a + \frac{\mathbb{1}(A=a)(Y-\mu_a)}{a\pi(x)+(1-a)\{1-\pi(x)\}}$. Then

$$\begin{aligned}
\mathbb{P}(\widehat{f}_1 - f_1) &= \mathbb{P}\left[\frac{A}{\widehat{\pi}(X)} \Big\{ Y - \widehat{\mu}_1(X) \Big\} + \Big\{ \widehat{\mu}_1(X) - \mu_1(X) \Big\} \right] \\
&= \mathbb{P}\left[\frac{\pi(X)}{\widehat{\pi}(X)} \Big\{ \mu_1(X) - \widehat{\mu}_1(X) \Big\} + \Big\{ \widehat{\mu}_1(X) - \mu_1(X) \Big\} \right] \\
&= \mathbb{P}\left[\frac{\pi(X) - \widehat{\pi}(X)}{\widehat{\pi}(X)} \Big\{ \mu_1(X) - \widehat{\mu}_1(X) \Big\} \right] \\
&\leq \mathbb{P}\left\{ \left| \frac{\pi(X) - \widehat{\pi}(X)}{\widehat{\pi}(X)} \right| \Big| \mu_1(X) - \widehat{\mu}_1(X) \Big| \right\} \\
&\leq \left(\frac{1}{\epsilon} \right) \mathbb{P}\left\{ \Big| \pi(X) - \widehat{\pi}(X) \Big| \Big| \mu_1(X) - \widehat{\mu}_1(X) \Big| \right\} \\
&\leq \left(\frac{1}{\epsilon} \right) \|\pi - \widehat{\pi}\| \|\mu_1 - \widehat{\mu}_1\| \\
&= O_{\mathbb{P}}(1/\sqrt{n}) O_{\mathbb{P}}(1/\sqrt{n}) = O_{\mathbb{P}}(1/n) = o_{\mathbb{P}}(1/\sqrt{n})
\end{aligned}$$

where the second and third lines used iterated expectation, the fifth used the bound on $\widehat{\pi}$, the sixth used Cauchy-Schwarz, and the last line used that $\widehat{\pi}$ and $\widehat{\mu}_a$ are root-n consistent due to the discrete (e.g., they can be represented as linear regression estimators, as mentioned above). The same exact logic follows for $\mathbb{P}(\widehat{f}_0 - f_0)$, which then yields the result since $T_1 + T_2 = o_{\mathbb{P}}(1/\sqrt{n})$. □

Theorem 4.1 shows that, when the covariates are discrete and low-dimensional, the causal effect estimators $\widehat{\psi}_{reg} = \widehat{\psi}_{ipw} = \widehat{\psi}_{dr}$ are all root-n consistent and asymptotically normal under only mild boundedness conditions. The key to proving this result was the analysis of the T_2 term; the logic used there will be repeated throughout the book going forward.

Theorem 4.1 gives confidence intervals (and thus hypothesis tests) as an immediate corollary.

Corollary 4.1. *Under the conditions of 4.1, an asymptotically valid confidence interval for the average treatment effect ψ is given by*

$$\widehat{\psi} \pm 1.96\sqrt{\widehat{var}(\widehat{f})/n}.$$

Remark 4.4. Although the regression, weighting, and doubly robust estimators are exactly equal, to construct confidence intervals we need to estimate the asymptotic variance with the empirical variance of the terms appearing in the doubly robust estimator.

In summary, when the measured covariates are sufficient to control confounding, and are discrete and low-dimensional, the choice of estimator is immaterial – regression, weighting, and doubly robust estimation are all numerically equivalent and efficient (note though that a simple difference-in-means estimator is no longer even consistent). In the next section, however, we will see that the story is much different in the more realistic scenario where the covariates are not all discrete, and so some modeling is necessary.

4.4.2 Regression & Matching

As mentioned earlier, the regression estimator (4.2) is perhaps most intuitive, since it immediately follows from plugging estimates into the identification result from Proposition 4.2. In practice, as in experiments with covariate adjustment, one could use simple parametric estimators (e.g., linear or logistic regression) or more flexible nonparametric estimators (e.g., kernel smoothing, random forests) of the regression functions μ_a.

First we will consider the case where μ_a is estimated with a finite-dimensional parametric model, i.e., it is assumed that

$$\mu_a(x) = \mu_a(x; \beta)$$

for some real-valued parameter $\beta \in \mathbb{R}^p$. A prominent example might include a logistic regression model $\mu_a(x; \beta) = \text{expit}(\beta_0 + \beta_1 a + \beta_2^{\mathrm{T}} x)$. The parameter β could be estimated via maximum likelihood or some relevant m-estimator, for example. Note that the discrete covariate setup can be viewed as a special case of this, since then for each $a = 0, 1$ the function $\mu_a(x)$ can be estimated with a saturated model with d parameters (e.g., $d - 1$ level indicators and an intercept).

In fact, we have already analyzed the estimator (4.2) in the case where it is assumed that μ_a follows a parametric model, in Theorem 3.1. Recall when we analyzed the parametric plug-in estimator we did not rely on the randomization at all, so the same analysis applies here. The relevant theorem is repeated below for posterity.

Theorem 4.2. *Let $f(x) = \mu_1(x) - \mu_0(x)$ and $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$, so that $\psi = \mathbb{E}\{f(X)\}$ is the average treatment effect. Assume the parametric model*

$$\mu_a(x) = \mu_a(x; \beta_a)$$

for some $\beta = (\beta_0, \beta_1) \in \mathbb{R}^p$. Suppose the estimator $\widehat{\beta} \in \mathbb{R}^p$ solves an estimating equation so that

$$\mathbb{P}_n\{m(Z; \widehat{\beta})\} = 0.$$

Assume $m(z; \beta) \in \mathbb{R}^p$ is Lipschitz in β, and that $\mathbb{E}\{m(z; \beta)\}$ is differentiable at the true β satisfying $\mathbb{E}\{m(Z; \beta)\} = 0$ with nonsingular derivative matrix. Then

$$\widehat{\psi}_{reg} - \psi = (\mathbb{P}_n - \mathbb{P}) g(Z; \beta) + o_{\mathbb{P}}(1/\sqrt{n})$$

where

$$g(z; \beta) = f(x; \beta) + \frac{\partial \mathbb{E}\{f(X; \beta)\}}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z; \beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(z; \beta)$$

and so is root-n consistent and asymptotically normal.

Theorem 3.1 shows that if a correct parametric model is available for the regression functions μ_a, and under some regularity conditions (essentially smoothness of the model), then the resulting plug-in regression estimator is root-n consistent and asymptotically normal. Confidence intervals can be constructed using an estimate of the closed-form asymptotic variance $\widehat{\text{var}}\{\widehat{g}(Z; \widehat{\beta})\}$, or via the bootstrap which is typically easier. Bootstrap estimates would be computed in the usual way: take a bootstrap sample (i.e., sample n observations with replacement), re-estimate the regression functions μ_a, and construct the plug-in estimator (and then repeat many times).

In practice, there is often not enough detailed background knowledge to justify a particular parametric model, especially in the presence of continuous covariates. This motivates the use of a nonparametric version of the regression estimator (4.2). However, in the nonparametric case, the analysis from Theorem 3.1 no longer applies, since μ_a is no longer assumed to be indexed by a finite-dimensional parameter. This means that the term $T_2 = \mathbb{P}(\widehat{f} - f)$ in the proof cannot be analyzed by simply differentiating with respect to β.

In fact, our discussion from Section 3.3.3 concerning nonparametric covariate adjustment in experiments applies as well, since as in the parametric case we did not rely on randomization there. To reiterate: when the regression functions μ_a are estimated nonparametrically, the term $T_2 = \mathbb{P}(\widehat{f} - f)$ cannot in general be expected to be $O_{\mathbb{P}}(1/\sqrt{n})$, so that the regression estimator would typically inherit slow convergence rates from estimating μ_a flexibly. This means that a nonparametric version of the regression estimator (4.2) has advantages over a parametric version because it will generally be consistent under weaker conditions, and thus more robust; however it has a potential disadvantage in that it will be less efficient, compared to if the parametric model is correctly specified. Further, it is not only slow rates that complicate the use of this estimator: for general nonparametric estimators it will often not have a tractable limiting distribution, and even when it does it would typically not be correctly centered (without undersmoothing), and so inference would be complicated at best.

Importantly, however, there are some exceptions to this story. Namely, if one specifies particular nonparametric estimators $\widehat{\mu}_a$ (e.g., based on kernels, splines, or nearest-neighbor regression), undersmooths (so that smaller bias of $\widehat{\mu}_a$ is traded off for larger variance, rather than the usual balancing), and makes some particular structural assumptions (e.g., that the regression functions are Hölder smooth), then nonparametric regression estimators can be root-n consistent (and sometimes asymptotically normal).

We will first discuss such an analysis of matching estimators, and then briefly mention series/spline estimators.

Matching

Matching is often treated as an entirely different adjustment approach, but it can be viewed as a particular nonparametric regression-based estimator, for example using k-nearest neighbor regression to estimate μ_a. Following Abadie and Imbens [2006], one simple version of matching uses $\widehat{\psi}_{reg}$ from (4.2) with

$$\widehat{\mu}_{A_i}(x_i) = Y_i \quad , \quad \widehat{\mu}_{1-A_i}(x_i) = \frac{1}{K} \sum_{j \in \mathcal{J}_K(i)} Y_j$$

where $\mathcal{J}_K(i)$ is the set of K indices with the treatment opposite of a_i, who are closest in terms of covariates, i.e., $\mathcal{J}_K(i) = \{j_1(i), ..., j_K(i)\}$ for

$$j_k(i) = \left\{ j : \sum_{\ell : A_\ell \neq A_i} \mathbb{1}\{\|X_\ell - X_i\|_2 \leq \|X_j - X_i\|_2\} = k \right\}$$

the index of the k^{th} closest match, for $\|\cdot\|_2$ the Euclidean norm. This corresponds to matching with replacement, since the same indices can reappear in $\mathcal{J}_K(i)$ for different subjects i. Of course, other schemes can also be used, e.g., based on matching without replacement, or via other distance metrics, or using a variable number of neighbors k, etc.

Abadie and Imbens [2006] show that, under usual iid sampling setup and consistency/exchangeability/positivity assumptions, and if

- the covariates $X \in \mathbb{R}^d$ are continuous, have compact and convex support, and density bounded away from zero, and if
- $\mu_a(x)$ is Lipschitz in x for $a \in \{0, 1\}$,

then for fixed K the matching estimator has (conditional) bias

$$\mathbb{P}_n \left((2A - 1) \Big[\mu_{1-A}(X) - \mathbb{E}\{\widehat{\mu}_{1-A}(X) \mid X, A\} \Big] \right) = O_{\mathbb{P}}(1/n^{1/d})$$

whose expected value is not of smaller order than $O(1/n^{2/d})$. This means that the estimator $\widehat{\psi}_{reg}$ when based on K-nearest neighbor matching is

- root-n consistent and asymptotically normal, if $d = 1$;
- root-n consistent but not necessarily asymptotically normal, if $d = 2$;
- not root-n consistent or asymptotically normal, if $d > 2$.

This illustrates the kind of result we discussed above heuristically: in general nonparametric versions of the regression estimator $\widehat{\psi}_{reg}$ are not expected to be root-n consistent and asymptotically normal, but some exceptions can occur, e.g., for particular K-nearest neighbor estimators, under smoothness conditions and/or low dimensions.

There are other examples of nonparametric estimators being root-n consistent with undersmoothing. Hahn [1998] has a nice paper that explores efficiency bounds for average treatment effects, and considers undersmoothed regression estimators based on series/spline regression. He shows that: if one uses particular orthonormal bases, and carefully tuned number of series terms, and if the propensity score and regression functions are infinitely differentiable, then a regression-type estimator is root-n consistent and asymptotically normal. Some caveats of these kinds of analyses are that in practice:

- it is difficult to construct bases satisfying the required regularity conditions;
- it can be even more difficult to pick the right number of basis terms (i.e., it requires undersmoothing – suboptimal tuning by trading less bias for more variance).

4.4.3 Weighting

The inverse-probability-weighted estimator (4.3) can be motivated from importance sampling or representativeness arguments, and depends on an estimate of the unknown propensity score $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$. In fact, its analysis is essentially the same as that of the regression-based estimator; the main difference is we define

$$\widehat{f} = \left\{ \frac{A}{\widehat{\pi}(X)} - \frac{1 - A}{1 - \widehat{\pi}(X)} \right\} Y$$

instead of $\widehat{f} = \widehat{\mu}_1 - \widehat{\mu}_0$.

First we will consider the case where the propensity score π is estimated with a parametric model, i.e.,

$$\widehat{\pi}(x) = \pi(x; \widehat{\beta}).$$

For example with logistic regression we might have $\pi(x; \beta) = \operatorname{expit}(\beta_0 + \beta_1^{\mathrm{T}} x)$. As with the regression estimator $\widehat{\psi}_{reg}$, the parameter β could be estimated via maximum likelihood or m-estimation. In the next result we give an analog of Theorem 3.1 for the parametric weighting estimator. Note that this setup can be viewed as a semiparametric model in which the propensity score π is restricted to follow a known parametric form, but the regression functions μ_a are left unrestricted.

Theorem 4.3. *Let* $f(z) = \left\{ \frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)} \right\} y$ *and* $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$*, so that* $\psi = \mathbb{E}\{f(Z)\}$ *is the average treatment effect. Assume the parametric model*

$$\pi(x) = \pi(x; \beta)$$

for some $\beta \in \mathbb{R}^p$*. Suppose the estimator* $\widehat{\beta} \in \mathbb{R}^p$ *solves an estimating equation so that*

$$\mathbb{P}_n\{m(Z; \widehat{\beta})\} = 0.$$

Assume $m(z; \beta) \in \mathbb{R}^p$ *is Lipschitz in* β*, and that* $\mathbb{E}\{m(z; \beta)\}$ *is differentiable at the true* β *satisfying* $\mathbb{E}\{m(Z; \beta)\} = 0$ *with nonsingular derivative matrix. Then*

$$\widehat{\psi}_{ipw} - \psi = (\mathbb{P}_n - \mathbb{P}) g(Z; \beta) + o_{\mathbb{P}}(1/\sqrt{n})$$

where

$$g(z; \beta) = f(z; \beta) + \frac{\partial \mathbb{E}\{f(Z; \beta)\}}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z; \beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(z; \beta)$$

and so is root-n consistent and asymptotically normal.

Proof. By Lemma 3.1 we have

$$\widehat{\psi} - \psi = Z^* + T_1 + T_2$$

where $Z^* = (\mathbb{P}_n - \mathbb{P})f$ and T_1 and T_2 defined accordingly. By Lemma 3.2 we have

$$\widehat{\beta} - \beta = (\mathbb{P}_n - \mathbb{P}) \left[\left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z;\beta) \right] + o_{\mathbb{P}}(1/\sqrt{n})$$

which also is enough to imply $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$. Further by the delta method we have

$$\begin{aligned} T_2 &= \mathbb{P}(\widehat{f} - f) = h(\widehat{\beta}) - h(\beta) \\ &= (\mathbb{P}_n - \mathbb{P}) \left[\frac{\partial h(\beta)}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z;\beta) \right] + o_{\mathbb{P}}(1/\sqrt{n}) \end{aligned}$$

for $h(\beta) = \mathbb{E}\{f(Z;\beta)\} = \mathbb{E}\left[\left\{ \frac{A}{\pi(X;\beta)} - \frac{1-A}{1-\pi(X;\beta)} \right\} Y \right]$. Combining terms gives the result. □

The analysis and interpretation of the weighting estimator parallels that of the regression estimator. In particular, Theorem 4.3 shows that if a correct parametric model is available for the propensity scores π, and under some regularity conditions (essentially smoothness of the model), then the resulting inverse-probability-weighted estimator is root-n consistent and asymptotically normal. Confidence intervals can be constructed using an estimate of the closed-form asymptotic variance $\widehat{\mathrm{var}}\{\widehat{g}(Z;\widehat{\beta})\}$, or via the bootstrap. Surprisingly, another variance estimator is available for the weighting estimator: namely, valid but conservative inference can be obtained via the variance estimate $\widehat{\mathrm{var}}\{f(Z;\widehat{\beta})\}$, which is the variance estimate that would be used if the propensity scores were known to be equal to $\widehat{\pi}(x) = \pi(x;\widehat{\beta})$ [Tsiatis, 2006]

As with regression estimoatrs, the analysis for weighting estimators also needs to be amended in the nonparametric case. Here, the term T_2 can essentially inherit the convergence rate of $\widehat{\pi}$ in the sense that

$$\begin{aligned} \mathbb{P}(\widehat{f} - f) &= \mathbb{P}\left\{ \left(\frac{\pi - \widehat{\pi}}{\widehat{\pi}} \right) \mu_1 - \left(\frac{\widehat{\pi} - \pi}{1 - \widehat{\pi}} \right) \mu_0 \right\} \\ &= \mathbb{P}\left\{ (\pi - \widehat{\pi}) \left(\frac{\mu_1}{\widehat{\pi}} + \frac{\mu_0}{1 - \widehat{\pi}} \right) \right\} \\ &\lesssim \mathbb{P}|\widehat{\pi} - \pi| \leq \sqrt{\mathbb{P}\{(\widehat{\pi} - \pi)^2\}} = \|\widehat{\pi} - \pi\| \end{aligned}$$

where the second inequality follows by Cauchy-Schwarz. Thus we should again not expect root-n rates for nonparametric weighting estimators, and confidence intervals (even using bootstrap) may not be correctly centered, and thus invalid.

However, again similar to the regression case, Hirano et al. [2003] show that undersmoothing can correct these issues, under some assumptions. In particular, they show that under some relatively strong smoothness conditions on the propensity score π (i.e., that π has seven times as many derivatives as it does dimensions d), and undersmoothing, a series-based nonparametric inverse-probability-weighted estimator is root-n consistent and asymptotically normal.

4.4.4 Doubly Robust Estimation

Although the doubly robust estimator (4.4) is perhaps less intuitive than the regression and inverse-probability-weighted estimators, it comes with important advantages. The classic advantage is that, when the nuisance functions π and μ_a are estimated with parametric models, the doubly robust estimator is root-n consistent and asymptotically normal even if one of the two models is completely misspecified. This classic result is given in the following theorem.

Theorem 4.4. *Let* $f(z) = \left\{\frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)}\right\}\{y - \mu_a(x)\} + \mu_1(x) - \mu_0(x)$, *so that* $\psi = \mathbb{E}\{f(Z)\}$ *is the average treatment effect. Assume at least one of the parametric models*

$$\pi(x) = \pi(x;\alpha) \ , \ \mu_a(x) = \mu_a(x;\beta_a)$$

for some $\eta = (\alpha, \beta_0, \beta_1) \in \mathbb{R}^p$. *Suppose the estimators* $\widehat{\eta} \in \mathbb{R}^p$ *solve an estimating equation so that*

$$\mathbb{P}_n\{m(Z;\widehat{\eta})\} = 0.$$

Assume $m(z;\eta) \in \mathbb{R}^p$ *is Lipschitz in* η, *and that* $\mathbb{E}\{m(z;\eta)\}$ *is differentiable at the true* η *satisfying* $\mathbb{E}\{m(Z;\eta)\} = 0$ *with nonsingular derivative matrix. Then*

$$\widehat{\psi}_{dr} - \psi = (\mathbb{P}_n - \mathbb{P})g(Z;\eta) + o_{\mathbb{P}}(1/\sqrt{n})$$

where

$$g(z;\eta) = f(z;\eta) + \frac{\partial \mathbb{E}\{f(Z;\eta)\}}{\partial \eta^{\mathrm{T}}}\left\{\frac{\partial \mathbb{E}(m(Z;\eta))}{\partial \eta^{\mathrm{T}}}\right\}^{-1} m(z;\eta)$$

and so is root-n consistent and asymptotically normal.

Proof. The proof is the same as Theorem 4.3, replacing f accordingly. □

Remark 4.5. When both the propensity score and regression models are correctly specified, it follows that

$$\widehat{\psi}_{dr} - \psi = (\mathbb{P}_n - \mathbb{P})f(Z) + o_{\mathbb{P}}(1/\sqrt{n})$$

since in that case $\frac{\partial \mathbb{E}\{f(Z;\eta)\}}{\partial \eta^{\mathrm{T}}} = 0$.

Theorem 4.4 shows the surprising phenomenon that the doubly robust estimator remains root-n consistent even if one of the two nuisance estimators is misspecified, giving the analyst "two chances" at fast rates and valid inference.

However, a skeptic might argue that the chance of correctly specifying a parametric model is essentially zero, in which case the chance of correctly specifying one of two parametric models is twice zero, and still zero. In fact the doubly robust estimator has an even more useful advantage. Namely, it can be root-n consistent and asymptotically normal even when the nuisance functions π and μ_a are estimated flexibly at slower than root-n rates, in a wide variety of settings. Surprisingly, this can be true even without

committing a priori to particular estimators or function classes, and without using any undersmoothing tricks. Contrast this with the estimators (4.2) and (4.3), which would generally not be root-n consistent when the nuisance functions are estimated at nonparametric rates, without requiring careful analysis of particular undersmoothed estimators.

Theorem 4.5. *Let* $f(z) = \left\{\frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)}\right\}\{y - \mu_a(x)\} + \mu_1(x) - \mu_0(x)$, *so that* $\psi = \mathbb{E}\{f(Z)\}$ *is the average treatment effect. Assume* $\|\widehat{f} - f\| = o_{\mathbb{P}}(1)$ *and either:*

1. f *and its estimate* $\widehat{f}$ *are contained in a Donsker class, or*
2. *the estimate* $\widehat{f}$ *is constructed from a separate independent sample.*

Also assume $\mathbb{P}(\widehat{\pi} \in [\epsilon, 1-\epsilon]) = 1$. *Then if* $\|\widehat{\pi} - \pi\| \sum_a \|\widehat{\mu}_a - \mu_a\| = o_{\mathbb{P}}(1/\sqrt{n})$ *it follows that*

$$\widehat{\psi}_{dr} - \psi = (\mathbb{P}_n - \mathbb{P})f(Z) + o_{\mathbb{P}}(1/\sqrt{n})$$

and so is root-n consistent and asymptotically normal.

Proof. By Lemma 3.1 we have

$$\widehat{\psi} - \psi = Z^* + T_1 + T_2$$

where $Z^* = (\mathbb{P}_n - \mathbb{P})f$ and T_1 and T_2 are defined accordingly. The consistency of $\widehat{f}$ in L_2 norm, together with either the Donsker condition or sample splitting, ensures that $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$. Finally for $T_2 = \mathbb{P}(\widehat{f} - f)$ and $f = f_1 - f_0$ we have

$$\begin{aligned}
\mathbb{P}(\widehat{f}_1 - f_1) &= \mathbb{P}\left\{\frac{A}{\widehat{\pi}}\Big(Y - \widehat{\mu}_A\Big) + \widehat{\mu}_1 - \mu_1\right\} \\
&= \mathbb{P}\left\{\left(\frac{\pi}{\widehat{\pi}} - 1\right)\Big(\mu_1 - \widehat{\mu}_1\Big)\right\} \\
&\leq \frac{1}{\epsilon}\mathbb{P}\left\{\Big|\pi - \widehat{\pi}\Big|\Big|\mu_1 - \widehat{\mu}_1\Big|\right\} \\
&\leq \frac{1}{\epsilon}\|\widehat{\pi} - \pi\|\|\widehat{\mu}_1 - \mu_1\|
\end{aligned}$$

where the second line follows by iterated expectation, the third by the bound on $\widehat{\pi}$, and the fourth by Cauchy-Schwarz. The exact same logic shows that $\mathbb{P}(\widehat{f}_0 - f_0) \leq \frac{1}{1-\epsilon}\|\widehat{\pi} - \pi\|\|\widehat{\mu}_0 - \mu_0\|$. Therefore $T_2 = o_{\mathbb{P}}(1/\sqrt{n})$ and the result follows. □

Remark 4.6. Note that Theorem 4.5 tells us that the doubly robust estimator is root-n consistent and asymptotically normal for a wide range of nuisance estimators $(\widehat{\pi}, \widehat{\mu}_a)$. For example, a sufficient condition for the product of L_2 norms being $o_{\mathbb{P}}(1/\sqrt{n})$ is that both $\widehat{\pi}$ and $\widehat{\mu}_a$ converge at (faster than) $n^{-1/4}$ rates. This could be satisfied for Hölder functions in $\mathcal{H}(s)$ if $s > d/2$, or for s-sparse functions if $s\log(d/s) < \sqrt{n}$. However, any product of the order $o_{\mathbb{P}}(1/\sqrt{n})$ would result in root-n consistency and asymptotic normality: for example if one nuisance function was estimated at root-n rates, the other would only need to be estimated consistently, at any rate.

Chapter 5

Instrumental Variables

5.0 Introduction

Thus far in non-experimental settings we have assumed there are no unmeasured confounders, i.e., that $A \perp\!\!\!\perp Y^a \mid X$ so that treatment is essentially randomized within levels of covariates. However, this may often not hold in observational studies. For example, some covariates may be too expensive or otherwise difficult or impossible to measure; or investigators may have simply missed some important unknown confounders, leaving them unmeasured.

Example 5.1. In Kennedy et al. [2019b] we studied infant mortality effects of medical treatment at high-level versus low-level NICUs, using observational data. We did not have all the pertinent information on each mother's health, e.g., detailed lab values and comorbidity information, to justify the no unmeasured confounding assumption.

Example 5.2. R.A. Fisher (a smoker) famously questioned whether smoking caused lung cancer. He posited several explanations of the large association between smoking and lung cancer, which involved unmeasured confounders.

- One theory was that lung cancer was preceded by an undiagnosed "chronic inflammation", which led people who had already started developing lung cancer to start smoking in order to alleviate symptoms. This would lead to higher cancer rates in smokers, not because smoking caused the cancer, but because early stages of the cancer led to smoking. Here the initial inflammation was an unmeasured confounder biasing study results.

- Another theory was that there was some latent genetic factor that led people to smoke more and also itself led to higher lung cancer rates.

Fisher's theories have since been refuted, with more careful epidemiological and even experimental lab studies (in the latter, where chemical components of tobacco were shown to produce tumors in mice, etc.). Fisher was a brilliant statistician, but his views on smoking have been relegated to the fringes, along with his similarly outdated views on eugenics, for example.

We saw in Section 4.1 that, without the no unmeasured confounding assumption, we can no longer reject a null of no treatment effect, even with infinite data. Without adding extra structure or assumptions, we can only bound the average treatment effect, and the bounds always include zero (and for binary outcomes always are of length one).

However, if you are lucky enough to find a special variable called an instrument, you can still make some progress towards causal inference.

5.0.1 Instrumental Variables

An instrumental variable Z must typically satisfy at least three conditions, which are displayed graphically in Figure 5.1.

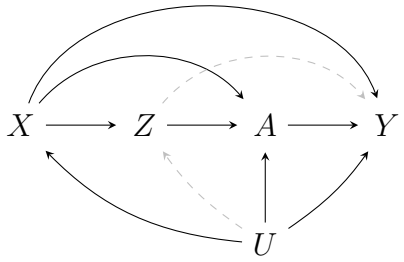


Figure 5.1: A directed acyclic graph representation of instrumental variable structure. Dashed arrows in light gray are assumed absent by instrumental variable assumptions.

More specifically, an instrument must satisfy at least:

1. Relevance: the instrument must be associated with treatment.
2. Exclusion Restriction: the instrument must not affect outcomes directly, only indirectly through treatment.
3. Unconfounded IV: the instrument must itself be unconfounded.

In Figure 5.1 relevance is conveyed by the presence of the path $Z \to A$, exclusion by the absence of the path $Z \to Y$, and unconfoundedness by the absence of the path $U \to Z$. Note that, in contrast to the previous chapter, the path $U \to A$ is allowed.

Remark 5.1. There is a lot of variation in the instrumental variables literature in terms of how the problem is framed and what particular assumptions are used. We will formalize several versions of the above assumptions shortly.

Many examples of instrumental variables can be found in both experiments and observational studies, across medicine, the social sciences, economics, and more.

Example 5.3. The archetypal example of an instrument is initial randomization (i.e., treatment assignment) in experiments with noncompliance. In this setup, the three foundational instrumental variable assumptions typically hold by design. Relevance holds since, generally, subjects are at least mildly persuaded to take their assigned treatment, on average, even if some do not comply. The exclusion restriction typically holds because, under double blinding, assignment should not affect outcomes directly since subjects and investigators are unaware of assignment. Unconfoundedness holds by design, since treatment *assignment* is randomized (unlike actual treatment receipt, which is often confounded by various factors).

Consider for example the Minneapolis Domestic Violence Experiment from the 1980s. In this experiment, among a pool of domestic violence offenders, police officers randomly selected some to be arrested versus receive other sanctions. Specifically, a third were randomized each to arrest, counseling, or separation from the partner. The goal of this experiment was to see if how these different penalties impact re-offense rates. However there was a relatively substantial amount of noncompliance; in particular, police arrested offenders more often than they were assigned to.

Example 5.4. Many prominent instrumental variables are distance measures, for example a subject's distance to nearest college, or health provider. For example, in Kennedy et al. [2019b] the instrument we posited was excess travel time to the nearest high-level versus low-level NICU. Thus someone with $Z = 20$ has to travel 20 extra minutes to get a high-level NICU, whereas someone with $Z = 0$ lives equally close to the nearest high-level and low-level NICU. In this case, relevance is likely to hold since living further away from a high-level NICU makes it more difficult to pursue, making the low-level NICU option more likely. The exclusion restriction may hold since the distance you live from the nearest high- versus low-level NICU is unlikely to directly affect outcomes. The unconfoundedness assumption is somewhat harder to justify, but one could argue it is less confounded than the treatment itself, at least conditional on socioeconomic status and other measured factors.

In another distance-based instrument example, Mauro et al. [2018] used distance from home to prison as an instrument to study effects of inmate visitation on recidivism.

Example 5.5. Preference measures (e.g., geographic regions, physician or provider preference) are also often used as instruments. For example, Brookhart et al. [2006] used physician's most recent NSAID prescription type as instrument to study effects of selective versus non-selective NSAIDs on GI complications. Another preference-based example is judge's sentencing severity; since cases are sometimes randomly assigned to judges, the unconfoundedness assumption can hold by design.

Example 5.6. Some other prominent examples of instruments include: calendar time (e.g., changes in recommendations or guidelines) and genes. The latter is called Mendelian

randomization and has been used to study effects of HDL cholesterol on heart disease, for example.

Remark 5.2. Note that the last few examples of instruments are more observational. In these cases the assumptions can sometimes be difficult to justify. In fact, often the hardest part of observational instrumental variable studies is finding and justifying a convincing instrument.

Remark 5.3. Instrumental variable methods have a long and interesting history. They originated in the 1920s and 1930s with studies of supply and demand by father and son team Philip and Sewall Wright. For many decades, identification of causal effects with instrumental variables relied heavily on parametric models. Further, the formal causal link was somewhat fuzzy early on, since there was not a well-defined language for counterfactuals. It was only later realized (by Imbens and Angrist [1994]) that there was a deeper nonparametric interpretation of these methods. This is a good lesson: there may be kernels of truth and insight even in seemingly simple restrictive methods.

5.1 Experiments with Noncompliance

5.1.1 Intention to Treat, As Treated, & Per Protocol Effects

Experiments are the easiest and most natural settings to think about instrumental variables. To focus discussion we will start there. Suppose n subjects are randomly *assigned* to treatment or control, denoted by $Z = 1$ and $Z = 0$, respectively. However, not all subjects comply with the assignment, so the actual treatment received A may be different from Z (in prior chapters on experiments we implicitly assumed $A = Z$). This means the actual treatment received A may be confounded, since it is not directly randomized. What can be done with the instrument Z?

Here the observed data are assumed to be an iid sample $(Z, A, Y) \sim \mathbb{P}$. In this setup one can define potential "outcomes" for both the treatment A and outcome Y when setting different values of the instrument:

- A^z is the treatment level that would have been observed if we had set $Z = z$.
- Y^a is the outcome that would have been observed if we had set $A = a$.
- Y^{za} is the outcome if both instrument and treatment were set to $(Z, A) = (z, a)$.
- $Y^z = Y^{zA^z}$ is the outcome that would have been observed if we only set $Z = z$.

Now the instrumental variable assumptions we discussed heuristically earlier can be formalized as follows.

1. Relevance: $\mathbb{P}(A^{z=1} = A^{z=0}) \neq 1$.

2. Exclusion Restriction: $Y^{za} = Y^a$.

3. Unconfounded IV: $Z \perp\!\!\!\perp (A^z, Y^z)$.

Again in an experiment these assumptions would typically hold by design.

There are three types of effects people typically consider in experiments with noncompliance:

$$\psi_{ITT} = \mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0) = \text{intention-to-treatment effect}$$

$$\psi_{AT} = \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0) = \text{as-treated effect}$$

$$\psi_{PP} = \mathbb{E}(Y \mid Z = A = 1) - \mathbb{E}(Y \mid Z = A = 0) = \text{per-protocol effect}$$

Remark 5.4. Per-protocol is sometimes used to refer to general effects of treatment received A, with intention-to-treat referring to effects of assignment Z.

There is good reason to prefer the ITT effect over the others. Specifically, under the first and third instrumental variable assumptions (even without the exclusion restriction), the ITT effect equals

$$\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0) = \mathbb{E}(Y^{z=1} - Y^{z=0})$$

i.e., the effect of assigning everyone versus no one to treatment. Note that this is not technically a "treatment" effect, but instead an effect of the instrument (which in this case is initial assignment). Therefore the ITT effect is a well-defined causal effect (not of A but of Z), whereas this is not generally true for the as-treated and per-protocol "effects", which are typically just non-causal associations.

For example, the as-treated effect is only the average treatment effect if actual treatment taken is completely randomized (i.e., $A \perp\!\!\!\perp Y^a$). This is typically unlikely: people stop taking treatment for many non-random reasons, e.g. maybe it is causing side effects, or maybe their symptoms improve so they feel treatment is unnecessary. One could collect relevant confounding covariates in the hope that $A \perp\!\!\!\perp Y^a \mid X$, but now the experiment is really just an observational study. Similarly, the per-protocol effect is only the average treatment effect if adherence $\mathbb{1}(A = Z)$ is completely randomized. Its bias is slightly more complicated.

For these reasons the ITT effect should generally be preferred over the other as-treated and per-protocol quantities. It is at least an effect of assignment, under just assumptions justified by the study design. The other two quantities are only effects in very restrictive and unusual settings.

In fact, under the exclusion restriction the ITT is also a kind of treatment effect since

$$\mathbb{E}(Y^{z=1} - Y^{z=0}) = \mathbb{E}(Y^{z=1,A^1} - Y^{z=0,A^0}) = \mathbb{E}(Y^{A^1} - Y^{A^0}).$$

This is the effect of an intervention that sets treatment to be the value it *would* take under $Z = 1$ versus $Z = 0$. This is an example of a so-called *stochastic intervention* effect, since the index of the potential outcome is a random variable, and not fixed or deterministic as for more standard parameters like the average effect $\mathbb{E}(Y^1 - Y^0)$.

Note also that the ITT effect is still defined even if relevance fails, i.e., if $A^{z=1} = A^{z=0}$ with probability one, but then

$$\mathbb{E}(Y^{A^1} - Y^{A^0}) = 0$$

so it equals zero.

Importantly, under the exclusion restriction the ITT effect can be used to test the sharp null hypothesis $H_0 : Y^1 = Y^0$ since

$$\begin{aligned}\mathbb{E}(Y^{z=1} - Y^{z=0}) &= \mathbb{E}(Y^{A^1} - Y^{A^0}) \\ &= \mathbb{E}\{(Y^1 - Y^0)\mathbb{1}(A^1 > A^0) + (Y^0 - Y^1)\mathbb{1}(A^1 < A^0)\} \\ &= \mathbb{E}[(Y^1 - Y^0)\{\mathbb{1}(A^1 > A^0) - \mathbb{1}(A^1 < A^0)\}]\end{aligned}$$

which equals zero if $Y^1 = Y^0$. Therefore testing $H_0 : ITT = 0$ provides a valid test of the sharp null of zero individual effect.

5.1.2 One-sided noncompliance

If there is one-sided noncompliance, i.e., if $Z = 0$ implies $A = 0$ (those assigned to control never take treatment), then it is possible to make some progress beyond the intention-to-treat effect, without adding any further assumptions.

One-sided noncompliance is quite common in practice, and especially in experiments. For example, in medical studies, control subjects may not have access to the drug under study if it is not yet on the market. In encouragement studies, some subjects assigned to receive encouragement may not be able to be contacted, but there would typically be no controls accidentally contacted, since this would require going beyond the study protocol. Examples of the latter include studies by: Green et al. (2003) exploring effects of canvassing on voter turnout, and Gerber et al. (2015) exploring effects of encouragement to vote among the formerly incarcerated. In the Minneapolis Domestic Violence Experiment, no one who was assigned to be arrested every received counseling (the only noncompliance occurred when police officers decided to arrest offenders assigned to not be arrested).

The next result shows that effects of removing treatment can be identified in experiments with one-sided noncompliance.

Theorem 5.1. *Let* $(Z, A, Y) \sim \mathbb{P}$*, and assume:*

1. *Relevance:* $\mathbb{P}(A^{z=1} = A^{z=0}) \neq 1$.
2. *Exclusion Restriction:* $Y^{za} = Y^a$.
3. *Unconfounded IV:* $Z \perp\!\!\!\perp (A^z, Y^z)$.

If there is one-sided noncompliance so that $A = 0$ *whenever* $Z = 0$*, then the effects of removing treatment in the whole population, and among the treated, are identified as*

$$\mathbb{E}(Y - Y^{a=0}) = \mathbb{E}(Y) - \mathbb{E}(Y \mid Z = 0)$$

and

$$\mathbb{E}(Y - Y^{a=0} \mid A = 1) = \frac{\mathbb{E}(Y) - \mathbb{E}(Y \mid Z = 0)}{\mathbb{P}(A = 1)}$$

respectively.

Proof. First note that, since $Z = 0 \implies A = 0$, it follows that $A^{z=0} = 0$. Therefore

$$\mathbb{E}(Y \mid Z = 0) = \mathbb{E}(Y^{z=0}) = \mathbb{E}(Y^{0,A^0}) = \mathbb{E}(Y^{A^0}) = \mathbb{E}(Y^{a=0})$$

where the first equality used unconfoundedness and consistency, the second consistency, the third the exclusion restriction, and the fourth the one-sided noncompliance. This gives the first result. For the effect on the treated, note that

$$\begin{aligned}\mathbb{E}(Y - Y^{a=0}) &= \mathbb{E}(Y - Y^{a=0} \mid A = 1)\mathbb{P}(A = 1) + \mathbb{E}(Y - Y^{a=0} \mid A = 0)\mathbb{P}(A = 0) \\ &= \mathbb{E}(Y - Y^{a=0} \mid A = 1)\mathbb{P}(A = 1)\end{aligned}$$

so the identification result follows from simply scaling by $\mathbb{P}(A = 1)$. □

One intuitive explanation of the identification result for $\mathbb{E}(Y^{a=0}) = \mathbb{E}(Y \mid Z = 0)$ is that, since Z is randomized and since $Z = 0$ implies $A = 0$, we get to see a random sample of $Y^{a=0}$ potential outcomes.

Just as in experiments with perfect compliance, in experiments with noncompliance where the unconfounded IV assumption holds marginally, covariate information could be useful to increase efficiency. And as before, in an observational study covariates may be needed to justify a conditional unconfounded IV assumption $A \perp\!\!\!\perp (A^z, Y^z) \mid X$ (however in observational studies, one-sided noncompliance is somewhat rare). The next result shows how effects of removing treatment can be identified in the presence of covariates.

Theorem 5.2. *Let* $(X, Z, A, Y) \sim \mathbb{P}$*, and assume:*

1. *Relevance:* $\mathbb{P}(A^{z=1} = A^{z=0}) \neq 1$.
2. *Exclusion Restriction:* $Y^{za} = Y^a$.

3. Unconfounded IV: $Z \perp\!\!\!\perp (A^z, Y^z) \mid X$.

If there is one-sided noncompliance so that $A = 0$ *whenever* $Z = 0$*, then the effects of removing treatment in the whole population, and among the treated, are identified as*

$$\mathbb{E}(Y - Y^{a=0}) = \mathbb{E}(Y) - \mathbb{E}\{\mathbb{E}(Y \mid X, Z = 0)\}$$

and

$$\mathbb{E}(Y - Y^{a=0} \mid A = 1) = \frac{\mathbb{E}(Y) - \mathbb{E}\{\mathbb{E}(Y \mid X, Z = 0)\}}{\mathbb{P}(A = 1)}$$

respectively.

Proof. The logic is the same as in the no-covariate case, with iterated expectation to go from conditional to marginal effects. □

For estimation of the above effects, we can use all the tools we learned from experiments and unconfounded observational studies. Both effects involve only marginal expectations and the quantity $\mathbb{E}\{\mathbb{E}(Y \mid X, Z = 0)\}$, which we studied in detail before (mathematically it makes no difference that Z plays the role that A played earlier). Therefore just as before, regression, weighting, and doubly robust estimators are all available.

The next result shows that if a root-n consistent and asymptotically normal estimator of $\mathbb{E}\{\mathbb{E}(Y \mid X, Z = 0)\}$ is available, then such an estimator is also available for the effect on the treated.

Proposition 5.1. *Assume* $\widehat{\theta}$ *is an estimator of* $\theta = \mathbb{E}\{\mathbb{E}(Y \mid X, Z = 0)\}$ *satisfying*

$$\widehat{\theta} - \theta = (\mathbb{P}_n - \mathbb{P})\phi + o_{\mathbb{P}}(1/\sqrt{n})$$

for some known function $\phi = \phi(Z; \mathbb{P})$*, i.e.,* $\widehat{\theta}$ *is root-n consistent and asymptotically normal. Then the estimator* $\widehat{\psi} = \frac{\mathbb{P}_n(Y) - \widehat{\theta}}{\mathbb{P}_n(A)}$ *of* $\psi = \mathbb{E}(Y - Y^{a=0} \mid A = 1)$ *satisfies*

$$\widehat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P}) \left\{ \frac{Y - \phi - \psi A}{\mathbb{E}(A)} \right\} + o_{\mathbb{P}}(1/\sqrt{n}).$$

Proof. Using the definitions of (ψ, θ) and their estimators, along with the asymptotic linearity of θ, we have

$$\begin{aligned}
\widehat{\psi} - \psi &= \frac{\mathbb{P}_n(Y) - \widehat{\theta}}{\mathbb{P}_n(A)} - \frac{\mathbb{E}(Y) - \theta}{\mathbb{E}(A)} \\
&= \frac{1}{\mathbb{P}_n(A)} \left\{ \mathbb{P}_n(Y) - \widehat{\theta} - \frac{\mathbb{P}_n(A)}{\mathbb{E}(A)} \mathbb{E}(Y - \theta) \right\} = \frac{1}{\mathbb{P}_n(A)} \Big\{ \mathbb{P}_n(Y - \phi - \psi A) \Big\} + o_{\mathbb{P}}(1/\sqrt{n}) \\
&= \frac{1}{\mathbb{E}(A)} \Big\{ \mathbb{P}_n(Y - \phi - \psi A) \Big\} + \left\{ \frac{1}{\mathbb{P}_n(A)} - \frac{1}{\mathbb{E}(A)} \right\} \Big\{ \mathbb{P}_n(Y - \phi - \psi A) \Big\} + o_{\mathbb{P}}(1/\sqrt{n})
\end{aligned}$$

$$= \frac{1}{\mathbb{E}(A)}\Big\{\mathbb{P}_n(Y - \phi - \psi A)\Big\} + o_{\mathbb{P}}(1/\sqrt{n})$$

where the last line follows since $\mathbb{E}(Y - \phi - \psi A) = \mathbb{E}(Y - \theta) - \psi\mathbb{E}(A) = 0$, so that $\mathbb{P}_n(Y - \phi - \psi A) = O_{\mathbb{P}}(1/\sqrt{n})$ and since $\mathbb{P}_n(A) - \mathbb{E}(A) = o_{\mathbb{P}}(1)$, so the product is $o_{\mathbb{P}}(1/\sqrt{n})$. □

Remark 5.5. An immediate implication of Proposition 5.1 is that asymptotically valid 95% confidence intervals can be constructed as

$$\widehat{\psi} \pm 1.96 \ \widehat{\text{sd}}\left\{\frac{Y - \widehat{\phi} - \widehat{\psi}A}{\mathbb{P}_n(A)}\right\}\Big/\sqrt{n}.$$

Remark 5.6. Recall that the root-n consistency and asymptotic normality of $\widehat{\theta}$ would hold for regression or weighting estimators in the discrete case, or when based on correct parametric models (or if nonparametrically estimated with particular undersmoothed estimators), and would hold for doubly robust estimators in the discrete case, or when based on correct parametric models (even if one model is misspecified), or if the nuisance functions are nonparametrically estimated as long as the estimators have L_2 error small enough that the product is $o_{\mathbb{P}}(1/\sqrt{n})$.

5.2 Classical IV Models

In many cases the noncompliance is not one-sided. Further the instrument Z and treatment A may not be binary. How does this change things? Can one still go beyond the ITT effect at all? Here we will discuss classical IV approaches, and then come to more modern extensions in the subsequent section.

The classical IV framework pre-1990 was largely based on the parametric two-stage model

$$A = \alpha_0 + \alpha_1 Z + \nu$$
$$Y = \beta_0 + \beta_1 A + \epsilon$$

for ν and ϵ some random error terms. Here the IV assumptions were typically framed as follows.

1. Relevance: $\text{cov}(A, Z) \neq 0$, i.e., $\alpha_1 \neq 0$.

2. Exclusion Restriction: Z is not included in second-stage Y equation.

3. Unconfounded IV: $\text{cov}(Z, \nu) = 0$ and $\text{cov}(Z, \epsilon) = 0$.

Alternatively, in the presence of covariates, one would have

$$A = \alpha_0 + \alpha_1 Z + \alpha_2^{\mathrm{T}} X + \nu$$
$$Y = \beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X + \epsilon$$

with $\mathrm{cov}(Z, \epsilon \mid X) = 0$.

Note that there are no potential outcomes in these models! How can this be causal? The answer is that these equations are meant to be of the "structural" variety we discussed in the first chapter. In other words, they indicate how nature actually assigns values, and are causal by assumption. Therefore, they can alternatively be written explicitly causally using potential outcomes as

$$A^z = \alpha_0 + \alpha_1 z + \nu$$
$$Y^{za} = \beta_0 + \beta_1 a + \epsilon$$

In the next result we show that the causal assumptions, together with the above explicitly causal model, imply the version of the model stated earlier.

Proposition 5.2. *The structural equation model*

$$A^z = \alpha_0 + \alpha_1 z + \nu$$
$$Y^{za} = \beta_0 + \beta_1 a + \epsilon$$

together with consistency and the three causal assumptions

1. *Relevance:* $\mathbb{P}(A^{z=1} = A^{z=0}) \neq 1$.
2. *Exclusion Restriction:* $Y^{za} = Y^a$.
3. *Unconfounded IV:* $Z \perp\!\!\!\perp (A^z, Y^z)$.

implies the corresponding observed data two-stage model and the three assumptions

1. *Relevance:* $cov(A, Z) \neq 0$, *i.e.,* $\alpha_1 \neq 0$.
2. *Exclusion Restriction:* Z *is not included in second-stage* Y *equation.*
3. *Unconfounded IV:* $cov(Z, \nu) = 0$ *and* $cov(Z, \epsilon) = 0$.

Proof. By consistency the causal model implies the observed data model

$$A = \alpha_0 + \alpha_1 Z + \nu$$
$$Y = \beta_0 + \beta_1 A + \epsilon.$$

The causal exclusion restriction immediately implies the classical version, and similarly with relevance. For the unconfounded IV assumption note

$$0 = \mathrm{cov}(Z, A^z) = \mathrm{cov}(Z, \alpha_0 + \alpha_1 z + \nu)$$

$$= \text{cov}(Z, \nu)$$

and similarly

$$\begin{aligned} 0 = \text{cov}(Z, Y^z) &= \text{cov}(Z, Y^{A^z}) \\ &= \text{cov}(Z, \beta_0 + \beta_1 A^z + \epsilon) \\ &= \beta_1 \text{cov}(Z, A^z) + \text{cov}(Z, \epsilon) \\ &= \text{cov}(Z, \epsilon) \end{aligned}$$

where the first equality follows by the unconfounded IV assumption, the second the exclusion restriction, the third the two-stage model, the fourth properties of covariance, and the last by the unconfounded IV assumption. □

One downside of structural equations of this form is that there is no explicit notation to indicate that they are causal rather than observed data equations. This can sometimes lead to ambiguity, and for this reason should really be paired with a graph (or written in potential outcome form above).

Another downside of these parametric structural equations is that they mix causal and statistical assumptions. For example, in the above two-stage models, there are causal assumptions being made simultaneously with linearity and additivity. This makes it hard to see whether identification comes from the causal or statistical parts of the model.

Remark 5.7. Under the structural assumptions above it follows that

$$\beta_1 = Y^{a=1} - Y^{a=0}.$$

In other words, β_1 is a subject-specific causal effect! However this only holds by fiat, e.g., by making the assumption that all subject-specific effects are the same. One could alternatively let the error ϵ vary under setting $a = 0, 1$, and then as long as it was mean-zero, it would only follow that $\beta_1 = \mathbb{E}(Y^{a=1} - Y^{a=0})$.

Note that, although we observe (Z, A, Y), we do not see the unknown parameters in the equations or the corresponding error terms. This begs the question of whether the effect β_1 is identified under the above model. The next result gives an answer.

Proposition 5.3. *Assume*

$$\begin{aligned} A &= \alpha_0 + \alpha_1 Z + \nu \\ Y &= \beta_0 + \beta_1 A + \epsilon \end{aligned}$$

Then if $cov(A, Z) \neq 0$ *and* $cov(Z, \epsilon) = 0$ *it follows that* β_1 *is identified as*

$$\beta_1 = \frac{cov(Y, Z)}{cov(A, Z)}.$$

Proof. We have

$$\begin{aligned}\text{cov}(Y, Z) &= \text{cov}(\beta_0 + \beta_1 A + \epsilon, Z) \\ &= \beta_1 \text{cov}(A, Z) + \text{cov}(\epsilon, Z) \\ &= \beta_1 \text{cov}(A, Z)\end{aligned}$$

where the second line used the $\text{cov}(Z, \epsilon) = 0$ assumption. The result follows by dividing both sides by $\text{cov}(A, Z) \neq 0$. □

Remark 5.8. The ratio in Proposition 5.3 is sometimes called the Wald estimand, after Wald (1940) used it in a measurement error problem.

The parameter β_1 is often estimated from data using the following two-stage least-squares procedure:

1. Do a least squares regression of A on Z, and get predicted values $\widehat{\alpha}_1 Z$.
2. Regress Y on predicted values $\widehat{\alpha}_1 Z$; then define $\widehat{\beta}_1$ as the coefficient estimate.

This is due to the fact that the parameter β_1 equals the population regression coefficient from regressing Y on $\alpha_1 Z$, as explained in the following proposition.

Proposition 5.4. *Consider the classical two-stage model with* $cov(A, Z) \neq 0$ *and* $cov(Z, \nu) = cov(Z, \epsilon) = 0$. *Then:*

1. β_1 *does not equal the coefficient in a population regression of* Y *on* A.
2. β_1 *equals the coefficient in a population regression of* Y *on the predicted value from a population regression of* A *on* Z.

Proof. For the first claim, note that the slope coefficient of a population regression of Y on A is

$$\begin{aligned}\frac{\text{cov}(A, Y)}{\text{var}(A)} &= \frac{\text{cov}(A, \beta_0 + \beta_1 A + \epsilon)}{\text{var}(A)} \\ &= \beta_1 + \frac{\text{cov}(A, \epsilon)}{\text{var}(A)}\end{aligned}$$

but $\text{cov}(A, \epsilon)$ can be non-zero. For the second, we have

$$\begin{aligned}\frac{\text{cov}(\alpha_1 Z, Y)}{\text{var}(\alpha_1 Z)} &= \frac{\text{cov}(\alpha_1 Z, \beta_0 + \beta_1 A + \epsilon)}{\text{var}(\alpha_1 Z)} \\ &= \beta_1 \frac{\alpha_1 \text{cov}(Z, A)}{\text{var}(\alpha_1 Z)}\end{aligned}$$

$$
\begin{aligned}
&= \beta_1 \frac{\alpha_1 \text{cov}(Z, \alpha_0 + \alpha_1 Z + \nu)}{\text{var}(\alpha_1 Z)} \\
&= \beta_1 \frac{\alpha_1^2 \text{cov}(Z, Z)}{\text{var}(\alpha_1 Z)} = \beta_1
\end{aligned}
$$

where the second line used $\text{cov}(Z, \epsilon) = 0$ and the last used $\text{cov}(Z, \nu) = 0$. □

When there are covariates, it is common to use the augmented two-stage model

$$
\begin{aligned}
A &= \alpha_0 + \alpha_1 Z + \alpha_2^{\mathrm{T}} X + \nu \\
Y &= \beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X + \epsilon
\end{aligned}
$$

Then by the same logic we have

$$
\begin{aligned}
\text{cov}(Z, Y \mid X) &= \text{cov}(Z, \beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X + \epsilon \mid X) \\
&= \beta_1 \text{cov}(A, Z \mid X)
\end{aligned}
$$

as long as $\text{cov}(Z, \epsilon \mid X) = 0$, so that

$$
\beta_1 = \frac{\text{cov}(Y, Z \mid X)}{\text{cov}(A, Z \mid X)} = \frac{\mathbb{E}\{\text{cov}(Y, Z \mid X)/\text{var}(Z \mid X)\}}{\mathbb{E}\{\text{cov}(A, Z \mid X)/\text{var}(Z \mid X)\}} \tag{5.1}
$$

and thus two-stage least-squares works the same way as before.

Remark 5.9. We can also write the IV estimand in (5.1) as

$$
\beta_1 = \frac{\mathbb{E}\{\mathbb{E}(Y \mid X, Z = 1) - \mathbb{E}(Y \mid X, Z = 0)\}}{\mathbb{E}\{\mathbb{E}(A \mid X, Z = 1) - \mathbb{E}(A \mid X, Z = 0)\}}
$$

Classical two-stage least-squares assumes the conditional quantities in the numerator and denominator are constant, not depending on X, but this is not strictly necessary.

In other words, above estimand equals the average treatment effect if the "first stage" of the model is unrestricted and the second is

$$
Y = h(X) + \beta_1 A + \epsilon
$$

for any h, since then

$$
\text{cov}(Y, Z \mid X) = \beta_1 \text{cov}(A, Z \mid X)
$$

This relaxation allows for some of the strong parametric assumptions to be avoided.

Let's take a step back. The classical IV model and two-stage least-squares methods are very popular, but they make some strong assumptions, for example: constant effects, additivity, linearity in all of X, Z, and A, as well as the errors. Further, the causal assumptions are framed within the modeling assumptions, making it hard to tease apart how much of the identification results are coming from the modeling versus

causal assumptions. These assumptions are inherently different, and often it can be much easier to reason about nonparametric assumptions like the exclusion restriction or unconfounded IV (e.g., based on whether certain confounders were measured or not, etc.). On the other hand, reasoning about linearity assumptions for latent variable regressions may prove very challenging; it is difficult to imagine what kind of substantive background knowledge could justify linearity or other distributional assumptions.

This raises a crucial question: what do these classical IV methods target if the strong assumptions are removed? Do they retain any meaning at all?

Imbens and Angrist [1994] brilliantly reverse-engineered these estimands (mostly focusing on the no covariate setting). This brings in an interesting reversal of the usual workflow: instead of defining a parameter of interest and assessing whether it is identified, here Imbens & Angrist (1994) took an identified statistical quantity, and asked what parameter it equaled under weaker assumptions. In other words, they tried to assess what if any causal quantity this method was getting at, without invoking the strong classical IV model assumptions.

5.3 Monotonicity & LATEs

The detective work of Imbens and Angrist [1994] led them to realize that the classical IV estimand equaled a particular subgroup effect, without using any of the parametric or constant effect assumptions, instead invoking the three base IV assumptions, along with a new assumption called monotonicity.

In words monotonicity says there are no subjects who always act *opposite* of what the instrument "encourages". More specifically, recall that with a binary instrument and treatment, every individual is of one of four types:

- compliers: $A^1 > A^0$
- always-takers: $A^1 = A^0 = 1$
- never-takers: $A^1 = A^0 = 0$
- defiers: $A^1 < A^0$

Compliers follow the encouragement of the instrument, in the sense that their treatment equals the instrument, i.e., $A = Z$, while always-takers and never-takers do not respond to the instrument. Defiers on the other hand take treatment when the instrument encourages control, and take control when the instrument encourages treatment. Monotonicity rules out the existence of this last group.

Assumption 5.1. Monotonicity: $\mathbb{P}(A^1 < A^0) = 0$.

Remark 5.10. The labeling of A and Z is arbitrary – for example we could have an instrument that encourages control rather than treatment, in which case compliers would have $A^1 < A^0$ and defiers $A^1 > A^0$. Without loss of generality we use the more intuitive labeling where $Z = 1$ encourages $A = 1$.

Remark 5.11. In addition to Imbens and Angrist [1994], monotonicity was also used by Robins [1989] and Manski [1990] outside of the instrumental variable setup; they considered bounding treatment effects under this and other assumptions.

Now we will state the fundamental result from Imbens and Angrist [1994], which says that the usual IV estimand actually equals a treatment effect among compliers, under the three base IV assumptions plus monotonicity.

Theorem 5.3. *Let* $(Z, A, Y) \sim \mathbb{P}$ *and assume:*

1. *Relevance:* $\mathbb{P}(A^{z=1} = A^{z=0}) \neq 1$.
2. *Exclusion Restriction:* $Y^{za} = Y^a$.
3. *Unconfounded IV:* $Z \perp\!\!\!\perp (A^z, Y^z)$.
4. *Monotonicity:* $\mathbb{P}(A^1 < A^0) = 0$.

along with consistency. Then

$$\frac{\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0)}{\mathbb{E}(A \mid Z = 1) - \mathbb{E}(A \mid Z = 0)} = \mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^{z=1} > A^{z=0}).$$

Proof. The numerator of the IV ratio estimand is

$$\begin{aligned}
\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0) &= \mathbb{E}(Y^{z=1} \mid Z = 1) - \mathbb{E}(Y^{z=0} \mid Z = 0) \\
&= \mathbb{E}(Y^{z=1} - Y^{z=0}) \\
&= \mathbb{E}(Y^{z=1,A^1} - Y^{z=0,A^0}) \\
&= \mathbb{E}(Y^{A^1} - Y^{A^0}) \\
&= \mathbb{E}[(Y^{a=1} - Y^{a=0})\{\mathbb{1}(A^1 > A^0) - \mathbb{1}(A^1 < A^0)\}]
\end{aligned}$$

where the first equality follows by consistency, the second unconfoundedness, the third consistency, the fourth the exclusion restriction, and the last by rearranging. Similarly the denominator of the IV estimand is

$$\begin{aligned}
\mathbb{E}(A \mid Z = 1) - \mathbb{E}(A \mid Z = 0) &= \mathbb{P}(A^1 = 1) - \mathbb{P}(A^0 = 1) \\
&= \mathbb{P}(A^1 > A^0) + \mathbb{P}(A^1 = A^0 = 1) - \mathbb{P}(A^1 < A^0) - \mathbb{P}(A^1 = A^0 = 1) \\
&= \mathbb{P}(A^1 > A^0) - \mathbb{P}(A^1 < A^0)
\end{aligned}$$

where the first equality follows by consistency and unconfoundedness, and the last since A is binary. Therefore without using monotonicity the IV estimand is

$$\psi = \frac{\mathbb{E}[(Y^{a=1} - Y^{a=0})\{\mathbb{1}(A^1 > A^0) - \mathbb{1}(A^1 < A^0)\}]}{\mathbb{P}(A^1 > A^0) - \mathbb{P}(A^1 > A^0)}.$$

Under monotonicity $\mathbb{1}(A^1 < A^0) = 0$ and $\mathbb{P}(A^1 > A^0) = 0$, so that

$$\psi = \frac{\mathbb{E}\{(Y^{a=1} - Y^{a=0})\mathbb{1}(A^1 > A^0)\}}{\mathbb{P}(A^1 > A^0)} = \mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^1 > A^0).$$

□

Remark 5.12. The subgroup effect

$$\mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^1 > A^0)$$

identified by the IV estimand is known as a *complier average treatment effect*, or also a *local average treatment effect* (LATE).

Before discussing the monotonicity assumption and the LATE parameter in more detail, note that introducing covariates into the picture presents no real additional complications from the identification perspective.

Proposition 5.5. *Let* $(X, Z, A, Y) \sim \mathbb{P}$ *and assume:*

1. *Relevance:* $\mathbb{P}(A^{z=1} = A^{z=0}) \neq 1$.
2. *Exclusion Restriction:* $Y^{za} = Y^a$.
3. *Unconfounded IV:* $Z \perp\!\!\!\perp (A^z, Y^z) \mid X$.
4. *Monotonicity:* $\mathbb{P}(A^1 < A^0) = 0$.

along with consistency and positivity $0 < \mathbb{P}(Z = 1 \mid X) < 1$ *with probability one. Then*

$$\mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^{z=1} > A^{z=0}) = \frac{\mathbb{E}\{\mathbb{E}(Y \mid X, Z = 1) - \mathbb{E}(Y \mid X, Z = 0)\}}{\mathbb{E}\{\mathbb{E}(A \mid X, Z = 1) - \mathbb{E}(A \mid X, Z = 0)\}}.$$

Proof. The logic is the same as in Theorem 5.3, arguing conditionally and then using iterated expectation. □

The monotonicity assumption and the LATE parameter have attracted a fair amount of controversy from some camps (see Imbens [2014] and Swanson and Hernán [2014] for recent debate). First consider monotonicity. Like most causal assumptions, monotonicity cannot be tested; in this case it is untestable because it depends on both

potential treatments A^z for $z = 0, 1$ (and as usual we only see one potential treatment $A = A^Z$). It is conceptually somewhat different from other assumptions we have used, in that it restricts the joint distribution of potential outcomes; in contrast the unconfoundedness assumption, for example, only restricts the marginal distribution of each potential outcome individually. However, it is similar to unconfoundedness in that it can hold by design: for example, if treatment is inaccessible to those assigned to control so that non-compliance is one-sided, then $A^0 = 0$ and monotonicity holds automatically.

Monotonicity can be a reasonable assumption, but there are of course examples where it would be violated. For example, consider preference-based instruments. It is quite possible for a defendant to be sentenced harshly with a less strict judge, but not with a more strict judge, e.g., if perhaps the generally more strict judge happens to have a soft spot for defendants of this type. Similarly, it is possible for a patient to be prescribed a drug even when seeing a physician who prescribes less, and not prescribed when seeing a physician who generally prescribes more. For example, maybe Doctor A prescribes more often overall but not for diabetics, while Doctor B prescribes less often overall but not for the physically active. Then someone who is diabetic and physically active might be prescribed under Doctor B but not Doctor A, contrary to the overall trend.

The LATE parameter is an interesting non-standard treatment effect, and has also led to some controversy. Note that it is an effect in a subgroup rather than the whole population; this is similar to, for example, effects on the treated subgroup, however it is different in that the subgroup here is not generally identified. In other words, we cannot directly observe who the compliers are, since only one of the potential outcomes A^1 or A^0 is observed, not both. On the one hand, under monotonicity the compliers are the only group for whom treatment effects make sense, since all others either always or never take treatment (under current conditions). Another justification for continuing to pursue complier effects is that they allow *something* causal to be learned in broken or "second-best" studies with unmeasured confounding, even without restricting effect heterogeneity [Imbens, 2010, 2014].

However, it may not be so useful from a policy perspective to only know the effect in an unidentifiable subgroup, especially if the compliers make up a small portion of the population. Robins and Greenland [1996] stressed early on that the complier subgroup is not identified, and gave examples where complier effects are not of primary policy interest. Pearl [2009b] says the complier "subpopulation cannot be identified and, more seriously, it cannot serve as a basis for policies." Deaton [2010] compares targeting local effects to the drunk who only looks for his keys near the lamppost, since that is where the light is. Swanson and Hernán [2014] state that complier effects "only pertain to an unknown subset of the population", and that "as we do not know who is a complier, we do not know to whom our new policy should apply."

However the subgroup of compliers are not entirely unknowable. Following Abadie [2003], Baiocchi et al. [2014] for example, the next result shows that it is possible to

identify and estimate the size of the complier population $\mathbb{P}(A^1 > A^0)$, as well as complier characteristics and covariate-specific chances of being a complier.

Proposition 5.6. *Let* $(X, Z, A, Y) \sim \mathbb{P}$ *and assume:*

1. *Relevance:* $\mathbb{P}(A^{z=1} = A^{z=0}) \neq 1$.
2. *Unconfounded IV:* $Z \perp\!\!\!\perp A^z \mid X$.
3. *Monotonicity:* $\mathbb{P}(A^1 < A^0) = 0$.

along with consistency and positivity $0 < \mathbb{P}(Z = 1 \mid X) < 1$ *with probability one. Then*

$$\begin{aligned}
\mathbb{P}(A^1 > A^0 \mid X) &= \mathbb{E}(A \mid X, Z = 1) - \mathbb{E}(A \mid X, Z = 0) \\
\mathbb{P}(A^1 > A^0) &= \mathbb{E}\{\mathbb{E}(A \mid X, Z = 1) - \mathbb{E}(A \mid X, Z = 0)\} \\
\mathbb{P}(V = v \mid A^1 > A^0) &= \frac{\mathbb{E}[\mathbb{1}(V = v)\{\mathbb{E}(A \mid X, Z = 1) - \mathbb{E}(A \mid X, Z = 0)\}]}{\mathbb{E}\{\mathbb{E}(A \mid X, Z = 1) - \mathbb{E}(A \mid X, Z = 0)\}}
\end{aligned}$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}(A \mid X, Z = 1) - \mathbb{E}(A \mid X, Z = 0) &= \mathbb{E}(A^1 - A^0 \mid X) \\
&= \mathbb{P}(A^1 > A^0 \mid X)
\end{aligned}$$

where the first equality follows by consistency, positivity, and unconfoundedness, and the second by monotonicity. The identifying expressions for strength $\mathbb{P}(A^1 > A^0)$ and $\mathbb{P}(V = v \mid A^1 > A^0)$ then follow by iterated expectation and Bayes' rule. □

Remark 5.13. Note that Proposition 5.6 required neither the exclusion restriction nor outcome unconfoundedness $Z \perp\!\!\!\perp Y^z \mid X$.

Remark 5.14. The fraction of the population that are compliers, denoted $\mathbb{P}(A^1 > A^0)$, is a fundamental property of an instrumental variable, often called the *strength* of the instrument. A strong instrument is one that is effective at encouraging subjects to take treatment, so that many have $A^1 = 1$ and $A^0 = 0$. A weak instrument is one that results in many always- and never-takers who are not responsive to the instrument.

In a similar vein, Kennedy et al. [2019a] recently showed that, in some settings, it is possible to accurately predict who compliers are, and obtain tight bounds on potentially more generalizable effects in identifiable subgroups. Instruments that yield more accurate complier predictions were termed *sharp* instruments.

5.4 Estimation of LATEs

In the previous section we showed that under some assumptions (specifically: consistency, positivity, relevance, the exclusion restriction, unconfoundedness for the IV, and monotonicity), the local average treatment effect is identified as

$$\psi = \frac{\mathbb{E}\{\mathbb{E}(Y \mid X, Z=1) - \mathbb{E}(Y \mid X, Z=0)\}}{\mathbb{E}\{\mathbb{E}(A \mid X, Z=1) - \mathbb{E}(A \mid X, Z=0)\}}.$$

At this point the goal becomes to estimate the statistical functional on the right-hand-side as well as possible. Note that it is a ratio of average treatment effects: i.e., an effect of Z on Y in the numerator and an effect of Z on A in the denominator. In fact our previous results immediately tell us how to estimate such quantities.

Specifically, let $\mu_z(X) = \mathbb{E}(Y \mid X, Z=z)$ denote the outcome regression, $\lambda_z(X) = \mathbb{E}(A \mid X, Z=z)$ the treatment regression, and $\pi(x) = \mathbb{P}(Z=1 \mid X=x)$ the instrument propensity score. Then the LATE can be estimated with regression, weighting, or both, via

$$\begin{aligned}
\widehat{\psi}_{reg} &= \frac{\mathbb{P}_n\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\}}{\mathbb{P}_n\{\widehat{\lambda}_1(X) - \widehat{\lambda}_0(X)\}} \\
\widehat{\psi}_{ipw} &= \frac{\mathbb{P}_n\left\{\frac{ZY}{\widehat{\pi}(X)} - \frac{(1-Z)Y}{1-\widehat{\pi}(X)}\right\}}{\mathbb{P}_n\left\{\frac{ZA}{\widehat{\pi}(X)} - \frac{(1-Z)A}{1-\widehat{\pi}(X)}\right\}} \\
\widehat{\psi}_{dr} &= \frac{\mathbb{P}_n\left[\left\{\frac{Z}{\widehat{\pi}(X)} - \frac{1-Z}{1-\widehat{\pi}(X)}\right\}\{Y - \widehat{\mu}_Z(X)\} + \widehat{\mu}_1(X) - \widehat{\mu}_0(X)\right]}{\mathbb{P}_n\left[\left\{\frac{Z}{\widehat{\pi}(X)} - \frac{1-Z}{1-\widehat{\pi}(X)}\right\}\{A - \widehat{\lambda}_Z(X)\} + \widehat{\lambda}_1(X) - \widehat{\lambda}_0(X)\right]}.
\end{aligned}$$

All of these estimators are ratios of estimators that were essentially already analyzed in previous chapters. First we state a result showing that if estimates of the numerator and denominator are root-n consistent and asymptotically normal, then so is the ratio.

Proposition 5.7. *Suppose* $\widehat{\theta}_t - \theta_t = \mathbb{P}_n(\phi_t) + o_{\mathbb{P}}(1/\sqrt{n})$ *for some mean-zero function* $\phi_t = \phi_t(Z; \mathbb{P})$ *and assume that* $\theta_2 \geq \epsilon > 0$. *Then*

$$\frac{\widehat{\theta}_1}{\widehat{\theta}_2} - \frac{\theta_1}{\theta_2} = \mathbb{P}_n\left\{\theta_2^{-1}\left(\phi_1 - \frac{\theta_1}{\theta_2}\phi_2\right)\right\} + o_{\mathbb{P}}(1/\sqrt{n})$$

Proof. Let $\psi = \theta_1/\theta_2$. Rearranging gives

$$\frac{\widehat{\theta}_1}{\widehat{\theta}_2} - \psi = \frac{1}{\widehat{\theta}_2}\left(\widehat{\theta}_1 - \psi\widehat{\theta}_2\right) = \frac{1}{\theta_2}\left(\widehat{\theta}_1 - \psi\widehat{\theta}_2\right) + \left(\frac{1}{\widehat{\theta}_2} - \frac{1}{\theta_2}\right)\left(\widehat{\theta}_1 - \psi\widehat{\theta}_2\right) \tag{5.2}$$

For the first term on the right-hand-side of (5.2) we have

$$\widehat{\theta}_1 - \psi\widehat{\theta}_2 = \mathbb{P}_n\phi_1 + \theta_1 - \psi\Big(\mathbb{P}_n\phi_2 + \theta_2\Big) + o_{\mathbb{P}}(1/\sqrt{n})$$

$$= \mathbb{P}_n(\phi_1 - \psi\phi_2) + o_{\mathbb{P}}(1/\sqrt{n})$$

where the first equality follows by definition of $\widehat{\theta}_t$ and the second since $\theta_1 - \psi\theta_2 = \theta_1 - \theta_1 = 0$.

And the second term on the right-hand-side of (5.2) is negligible since

$$\begin{aligned}\left(\frac{1}{\widehat{\theta}_2} - \frac{1}{\theta_2}\right)\left(\widehat{\theta}_1 - \psi\widehat{\theta}_2\right) &= \left(\frac{1}{\widehat{\theta}_2} - \frac{1}{\theta_2}\right)\left(\widehat{\theta}_1 - \psi\widehat{\theta}_2\right) \\ &= \left(\frac{\theta_2 - \widehat{\theta}_2}{\theta_2\widehat{\theta}_2}\right)\left(\widehat{\theta}_1 - \psi\widehat{\theta}_2\right) \\ &= O_{\mathbb{P}}(1/\sqrt{n})O_{\mathbb{P}}(1/\sqrt{n}) \\ &= O_{\mathbb{P}}(1/n) = o_{\mathbb{P}}(1/\sqrt{n})\end{aligned}$$

where the third equality used the fact that $\widehat{\theta}_2$ and $\widehat{\theta}_1 - \psi\widehat{\theta}_2$ are root-n consistent (and that the latter is an estimator of zero). This gives the result. □

Proposition 5.7 shows that root-n consistent and asymptotically normal estimation of the numerator and denominator, separately, is enough to ensure root-n consistency and asymptotic normality of the ratio estimator itself. In earlier chapters we studied exactly when this should be expected for regression, weighting, and doubly robust estimators of these quantities. In what follows we give a brief summary of these results (for simplicity we assume the estimators use sample splitting to separate $\mathbb{P}_n$ from the training data on which the nuisance estimators $(\widehat{\pi}, \widehat{\lambda}, \widehat{\mu})$ are built).

When $\widehat{\pi} = \pi$ is known, as in an experiment:

- The weighting estimator $\widehat{\psi}_{ipw}$ is root-n consistent and asymptotically normal.
- The regression estimator $\widehat{\psi}_{reg}$ would be root-n consistent and asymptotically normal if the estimators $(\widehat{\lambda}, \widehat{\mu})$ are built from correctly specified parametric models (though the models could be misspecified if π is constant and the relevant score equations are solved, e.g., by including an intercept and main effect for treatment).
- The doubly robust estimator $\widehat{\psi}_{reg}$ is root-n consistent and asymptotically normal as long as the estimators $(\widehat{\lambda}, \widehat{\mu})$ converge to some fixed function in L_2 norm, at any rate, and optimally efficient if $(\widehat{\lambda}, \widehat{\mu})$ are consistent in L_2 norm, at any rate.

When π is unknown but X is discrete and of fixed dimension (and if the nuisance estimators are built using the empirical distribution):

- All three estimators are numerically equivalent, root-n consistent, asymptotically normal, and optimally efficient.

When π is unknown and X is high-dimensional and/or contains continuous components (so the nuisance estimators require some smoothing):

- The regression and weighting estimators would be root-n consistent and asymptotically normal when the relevant nuisance functions are (a) estimated with correct parametric models, or (b) estimated nonparametrically but with undersmoothing and particular e.g., smoothness assumptions.

- The doubly robust estimator is root-n consistent and asymptotically normal when (a) the nuisance functions are estimated with parametric models, and either the models for π or (λ, μ) (or both) are correct, or (b) if the nuisance functions are estimated consistently with errors satisfying $\|\widehat{\pi}-\pi\|(\|\widehat{\lambda}-\lambda\|+\|\widehat{\mu}-\mu\|) = o_{\mathbb{P}}(1/\sqrt{n})$ (for example, a sufficient condition is if all are estimated at $n^{-1/4}$ rates).

The next result gives sufficient conditions for the doubly robust estimator to be root-n consistent and asymptotically normal.

Theorem 5.4. *Let*

$$f_1(z) = \left\{\frac{z}{\pi(x)} - \frac{1-z}{1-\pi(x)}\right\}\{y - \mu_z(x)\} + \mu_1(x) - \mu_0(x)$$
$$f_2(z) = \left\{\frac{z}{\pi(x)} - \frac{1-z}{1-\pi(x)}\right\}\{a - \lambda_z(x)\} + \lambda_1(x) - \lambda_0(x)$$

so that $\psi = \mathbb{E}\{f_1(Z)/f_2(Z)\}$ *is the local average treatment effect. Assume* $\|\widehat{f}_j - f_j\| = o_{\mathbb{P}}(1)$ *and either:*

1. f *and its estimate* $\widehat{f}$ *are contained in a Donsker class, or*

2. *the estimate* $\widehat{f}$ *is constructed from a separate independent sample.*

Also assume $\mathbb{P}(\widehat{\pi} \in [\epsilon, 1-\epsilon]) = 1$ *and* $\mathbb{E}(f_2) \geq \epsilon > 0$. *Then if* $\|\widehat{\pi} - \pi\| \sum_z (\|\widehat{\lambda}_z - \lambda_z\| + \|\widehat{\mu}_z - \mu_z\|) = o_{\mathbb{P}}(1/\sqrt{n})$ *it follows that*

$$\widehat{\psi}_{dr} - \psi = (\mathbb{P}_n - \mathbb{P})\left[\frac{f_1(Z) - \psi f_2(Z)}{\mathbb{E}\{f_2(Z)\}}\right] + o_{\mathbb{P}}(1/\sqrt{n})$$

and so is root-n consistent and asymptotically normal.

Proof. Let $\widehat{\theta}_j = \mathbb{P}_n\{\widehat{f}_j(Z)\}$. Then the assumptions of Theorem 4.5 are satisfied and we have

$$\widehat{\theta}_j - \theta_j = (\mathbb{P}_n - \mathbb{P})f_j(Z) + o_{\mathbb{P}}(1/\sqrt{n})$$

so that Proposition 5.7 gives the result. □

5.5 Bounds on ATE

The controversy surrounding LATEs is often based on the fact that it is an effect in a non-identified subgroup, rather than the plausibility of monotonicity. (Though sometimes it is the plausibility of monotonicity that is problematic, and also recall that in some cases it is possible to precisely identify who the compliers are). Suppose one was set on estimating the average treatment effect in the population; of course this parameter is not identifed. One solution then is to add classical IV model assumptions, but these are somewhat fragile and often hard to justify (this is similar to equating the LATE with the ATE merely by assumption). An alternative approach would be to live with the fact that the average effect is not point identified, and instead try to bound its possible values. Such bounds are given in the following theorem.

Theorem 5.5. *Let $(Z, A, Y) \sim \mathbb{P}$ with $Y \in [0,1]$ and assume consistency, positivity, the exclusion restriction, and monotonicity. Then*

$$\psi_\ell \leq \mathbb{E}(Y^{a=1} - Y^{a=0}) \leq \psi_u$$

for

$$\psi_\ell = \mathbb{E}(AY \mid Z=1) - \mathbb{E}\{Y(1-A)+A \mid Z=0\}\,,\ \psi_u = \psi_\ell + 1 - \mathbb{E}(A \mid Z=1) + \mathbb{E}(A \mid Z=0)$$

Proof. Note that

$$\begin{aligned}\mathbb{E}(Y^{a=1} - Y^{a=0}) &= \mathbb{E}\{(Y^{a=1} - Y^{a=0})\mathbb{1}(A^1 > A^0)\} \\ &\quad + \sum_{t=0}^{1} \mathbb{E}\{(Y^{a=1} - Y^{a=0})\mathbb{1}(A^0 = A^1 = t)\} \qquad (5.3)\end{aligned}$$

The first term is identified as the numerator of the IV estimand, i.e., the intention-to-treat effect $\psi_{itt} = \mathbb{E}(Y \mid Z=1) - \mathbb{E}(Y \mid Z=0)$. For the second term, we have

$$\begin{aligned}\mathbb{E}\{Y^{a=1}\mathbb{1}(A^{z=0} = A^{z=1} = 1)\} &= \mathbb{E}(Y^{a=1}A^{z=0}) = \mathbb{E}(Y^{a=A^0}A^{z=0}) \\ &= \mathbb{E}(Y^{z=0}A^{z=0}) = \mathbb{E}(YA \mid Z=0)\end{aligned}$$

where the first two equalities follow by monotonicity and consistency, the third by the exclusion restriction and consistency, and the last by IV unconfoundedness. Similarly

$$\begin{aligned}\mathbb{E}\{Y^{a=0}\mathbb{1}(A^{z=0} = A^{z=1} = 0)\} &= \mathbb{E}\{Y^{a=0}\mathbb{1}(A^{z=1} = 0)\} = \mathbb{E}\{Y^{a=A^1}\mathbb{1}(A^{z=1} = 0)\} \\ &= \mathbb{E}\{Y^{z=1}\mathbb{1}(A^{z=1} = 0)\} = \mathbb{E}\{Y(1-A) \mid Z=1\}\end{aligned}$$

Therefore using the fact that $\mathbb{P}(Y \in [0,1]) = 1$ and plugging in corresponding upper and lower bounds to (5.3) gives

$$\begin{aligned}\psi_{itt} + 0 - \mathbb{E}(Y(1-A) \mid Z=1) + \mathbb{E}(YA \mid Z=0) - 1 \times \mathbb{E}(A \mid Z=0) &\leq \psi_{ate} \\ \leq \psi_{itt} + 1 \times \mathbb{E}(1-A \mid Z=1) - \mathbb{E}(Y(1-A) \mid Z=1) + \mathbb{E}(YA \mid Z=0) - 0\end{aligned}$$

□

Remark 5.15. Under the IV assumptions, the length of the bounds in Theorem 5.5 is exactly equal to the proportion of non-compliers $\mathbb{P}(A^1 < A^0)$.

Appendix A

Notation Guide

Y^a	Potential outcome under treatment/exposure $A = a$
$\perp\!\!\!\perp$	Statistically independent
$\xrightarrow{p}$	Convergence in probability
$\rightsquigarrow$	Convergence in distribution
$O_{\mathbb{P}}(1)$	Bounded in probability
$o_{\mathbb{P}}(1)$	Converging in probability to zero
$\mathbb{P}_n$	Sample average operator, as in $\mathbb{P}_n(\widehat{f}) = \mathbb{P}_n\{\widehat{f}(Z)\} = \frac{1}{n}\sum_{i=1}^{n}\widehat{f}(Z_i)$
$\mathbb{P}$	Conditional expectation given the sample operator, as in $\mathbb{P}(\widehat{f}) = \int \widehat{f}(z)\, d\mathbb{P}(z)$
$\|\cdot\|$	$L_2(\mathbb{P})$ norm $\|f\| = \sqrt{\mathbb{P}(f^2)}$ or Euclidean norm, depending on context
$\|\cdot\|_1$	$L_1(\mathbb{P})$ norm $\|f\|_1 = \mathbb{P}(\|f\|)$
$\|\cdot\|_\infty$	L_∞ or supremum norm $\|f\|_\infty = \sup_z \|f(z)\|$
$\mathcal{H}(s)$	Hölder class of functions with smoothness index s
$\lesssim$	Less than or equal, up to a constant multiplier

Bibliography

A. Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

M. Baiocchi, J. Cheng, and D. S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press, 1993.

D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013.

M. A. Brookhart, P. Wang, D. H. Solomon, and S. Schneeweiss. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology*, 17(3):268, 2006.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.

M. Davidian, A. A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science*, 20(3):261, 2005.

A. Deaton. Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2):424–455, 2010.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, pages 237–249, 2008.

S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, pages 29–46, 1999.

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G. W. Imbens. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature*, 48(2):399–423, 2010.

G. W. Imbens. Instrumental variables: An econometrician's perspective (with discussion). *Statistical Science*, 29(3):323–358, 2014.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245, 2017.

E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics (to appear)*, 2019a.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019b.

S. Leon, A. A. Tsiatis, and M. Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055, 2003.

C. F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

J. A. Mauro, E. H. Kennedy, and D. Nagin. Instrumental variable methods using dynamic interventions. *Journal of the Royal Statistical Society, Series A*, 2018.

D. Michaels. *Doubt is their product: how industry's assault on science threatens your health*. Oxford University Press, 2008.

J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009a.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009b.

J. Pfanzagl. *Contributions to a general asymptotic statistical theory*, volume 13. Springer, 1982.

J. M. Robins. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health Service Research Methodology: A Focus on AIDS*, pages 113–159, 1989.

J. M. Robins and S. Greenland. Comment: Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):456–458, 1996.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995.

J. M. Robins and A. Rotnitzky. Comments on: Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:920–936, 2001.

J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87(1): 113–124, 2000.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

D. B. Rubin and M. J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4(1), 2008.

S. A. Swanson and M. A. Hernán. Think globally, act globally: An epidemiologist's perspective on instrumental variable estimation. *Statistical Science*, 29(3):371–374, 2014.

Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer, 2003.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.

L. Yang and A. A. Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715, 2008.